

Wireless Throughput and Energy Efficiency with Random Arrivals and Statistical Queueing Constraints

Mustafa Ozmen and M. Cenk Gursoy

Abstract—Throughput and energy efficiency in fading channels are studied in the presence of randomly arriving data and statistical queueing constraints. In particular, Markovian arrival models including discrete-time Markov, Markov fluid, and Markov-modulated Poisson sources are considered. Employing the effective bandwidth of time-varying sources and effective capacity of time-varying wireless transmissions, maximum average arrival rates in the presence of statistical queueing constraints are characterized. For the two-state (ON/OFF) source models, throughput is determined in closed-form as a function of the source statistics, channel characteristics, and quality of service (QoS) constraints. Throughput is further studied in certain asymptotic regimes. Furthermore, energy efficiency is analyzed by determining the minimum energy per bit and wideband slope in the low signal-to-noise ratio (SNR) regime. Overall, the impact of source characteristics, QoS requirements, and channel fading correlations on the throughput and energy efficiency of wireless systems is identified.

Index Terms—effective bandwidth, effective capacity, energy efficiency, fading channels, Markovian source models, maximum average arrival rates, minimum energy per bit, QoS provisioning, wideband slope, wireless throughput.

I. INTRODUCTION

A. Motivation

Mobile data traffic has experienced unprecedented growth recently and is predicted to grow even further over the coming years. For instance, it is projected that global mobile data traffic, which already grew 81 percent in 2013, will increase 11-fold between 2013 and 2018, reaching 15.9 exabytes per month by 2018 [1]. As one of the main driving forces behind this growth, wireless transmission of multimedia content has significantly increased in volume and is expected to be the dominant traffic in data communications. Indeed, mobile video traffic was already 53 percent of the traffic by the end of 2013 and is predicted to increase 14-fold between 2013 and 2018, accounting for over two-thirds of the world's mobile data traffic by 2018 [1].

This exponential growth in the flow of mobile data and multimedia content has significant implications on wireless networks. For one, wireless multimedia traffic requires certain quality-of-service (QoS) guarantees. For instance, in voice over IP (VoIP), multimedia streaming, interactive video, and online gaming applications, constraints on delay, packet loss, or buffer overflow probabilities need to be imposed so that acceptable performance and quality levels can be met for the end-users. Another consequence is heterogeneity in network traffic. Wireless networks now carry

heterogeneous traffic in diverse environments, and successful design of networks, efficient use of resources, and effective QoS provisioning for multimedia communications critically depend on the appropriate choice of source traffic models. For instance, while voice traffic can be accurately modeled as an ON/OFF process with fixed-rate data arrivals in the ON state, data traffic can be bursty and video traffic, which exhibits correlations, can be modeled statistically using autoregressive, Markovian, or Markov-modulated processes [2].

Finally, it is important to note that this increased traffic together with the given QoS requirements need to be supported by wireless systems equipped with only limited bandwidth and power resources. Especially, due to limited energy available for mobile units and rising energy costs and environmental concerns, energy efficiency in wireless communications is a key concern (see e.g., [3] and [4]). Therefore, it is crucial to identify the fundamental performance limits (e.g., in terms of maximum achievable throughput and minimum energy per bit) in order to determine how to most effectively utilize the scarce resources. With this motivation, in this paper we investigate the throughput and energy efficiency in fading channels when data arrivals are in general random, and QoS constraints in the form of limitations on the asymptotic buffer overflow probabilities are imposed.

B. Literature Overview

Satisfying QoS requirements is critical for most communication networks, and how to satisfy QoS constraints for various source traffic models has been one of the key considerations in the networking literature. In particular, besides conventional queueing theory, network calculus has been introduced by Cruz in early 1990s as a theory to address the delay and other deterministic service guarantees in networks by dealing with queueing systems [5] – [7]. Subsequently, Chang in [8] developed the stochastic version of the network calculus. More specifically, the theory of effective bandwidth of a time-varying source has been formulated to identify the minimum amount of transmission rate that is needed to satisfy the statistical QoS requirements (see also [9] – [13]). This theory is based on the logarithmic moment generating function of the arrival process and is related to the large deviation principle. Moreover, statistical QoS constraints are imposed as limitations on buffer/delay violation probabilities. Effective bandwidths of various source models have been investigated extensively in the literature. For instance, Elwalid and Mitra studied the effective bandwidth of Markovian traffic sources (including Markov-modulated fluid and Markov-modulated Poisson sources) in [14] under constraints on the buffer overflow probability. It is shown that effective bandwidth

The authors are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244. (e-mails: mozman@syr.edu, mcgursoy@syr.edu)

The material in this paper was presented in part at the 2012 IEEE International Symposium on Information Theory.

is given by the maximum eigenvalue of a matrix derived from source parameters and service requirements. In [15], effective bandwidth formulations were provided for multi-class Markov fluids as well as memoryless (Poisson) and discrete-time Markov sources. In [16], the authors studied the effective bandwidths of general stationary sources and derived a first order approximation of the effective bandwidth in terms of the mean arrival rate and index of dispersion.

In wireless communications, the instantaneous channel capacity varies randomly depending on the channel conditions. Hence, in addition to the source characteristics, transmission rates for reliable communication are also time-varying. In such cases, randomly time-varying servers can be considered in the queueing system model. Indeed, motivated by the wireless channel, Stolyar in [17], Venkataramanan and Lin in [18], and Sadiq and de Veciana in [19] employed tools from the theory of large deviations and investigated scheduling rules (e.g., MaxWeight, Exponential, and Radial Sum-Rate Monotonic scheduling) while controlling the large deviations of queues. Following another method, the time-varying channel capacity can be incorporated into the theory of effective bandwidth by regarding the channel service process as a time-varying source with negative rate and using the source multiplexing rule ([11, Example 9.2.2]). Using a similar approach, as a dual concept to effective bandwidth, Wu and Negi defined in [20] the effective capacity, which describes the maximum constant arrival rate that a given time-varying service process can support while satisfying the statistical QoS requirements. Indeed, work in [20] revitalized the consideration of statistical queueing constraints in the context of wireless communications, and the effective capacity of wireless transmissions has been investigated intensively in various settings (see e.g., [21]–[35]). For instance, Tang and Zhang in [22] considered the effective capacity when both the receiver and transmitter know the instantaneous channel gains, and derived the optimal power policy that maximizes the system throughput under QoS constraints. Liu *et al.* in [23] considered fixed-rate transmission schemes and analyzed the effective capacity and related resource requirements for Markov wireless channel models and Markov fluid sources. In [26] and [27], effective capacity of cognitive radio channels was studied. In [28], multi-antenna communication in the presence of queueing limitations was investigated. Soret *et al.* in [30] addressed correlated Rayleigh fading channels and studied the effective capacity under different adaptive rate policies. In this study, performance in the presence of probabilistic delay constraints and variable rate sources was also analyzed by considering a Gaussian autoregressive source model. Energy efficiency in the presence of QoS limitations was addressed in [32] – [35].

C. Contributions

We note that the studies on the effective capacity of wireless channels have primarily concentrated on constant arrival rates in

the analysis of the throughput and energy efficiency¹. Departing from this approach, we in this paper explicitly take into account the randomness and burstiness of the source traffic. In particular, we address Markovian source models including discrete-time Markov, Markov fluid, and Markov modulated Poisson sources, and conduct a performance analysis. More specifically, our contributions can be listed as follows:

- A framework with which source randomness can be incorporated in the throughput analysis of wireless transmissions is provided.
- For two-state (ON/OFF) source models, closed-form expressions are obtained for the maximum average arrival rate in terms of the source statistics, effective capacity of wireless transmissions, and the QoS exponent θ , which quantifies how strict the QoS constraints are.
- Throughput is characterized in the low- θ and high-SNR regimes.
- An energy efficiency analysis is conducted and minimum energy per bit and wideband slope expressions are determined for both constant and random arrival models.
- Via both analytical and numerical results, the impact of source randomness, fading correlations, and queueing constraints on the wireless throughput and energy efficiency is identified.

The remainder of this paper is organized as follows. In Section II-A, we describe the channel model. Sections II-B and II-C contain the preliminaries regarding the statistical queueing constraints, effective bandwidth, and effective capacity. In Section III, we provide our characterizations of the throughput with Markovian source models by analyzing the maximum average arrival rates. We conduct an energy efficiency analysis in Section IV for both constant and Markovian arrivals. Finally, concluding remarks are given in Section V. Proofs are relegated to the Appendix.

II. SYSTEM MODEL

A. Channel Model

As depicted in Fig. 1, we consider a point-to-point link with a single transmitter and single receiver. In this system, the data generated by the source is initially stored in a buffer at the transmitter before it is transmitted over a wireless channel. We consider a flat-fading channel between the transmitter and receiver, and assume a block-fading model with a block duration of m symbols. Hence, fading varies independently from one block to another. On the other hand, we further assume that in each block duration of m symbols, fading can be arbitrarily correlated. The channel input-output relation within each block

¹To the best of our knowledge, the two exceptions to this are references [23] and [30] as also described in Section I-B. However, these studies have different modeling assumptions for the sources and/or wireless transmissions from what we have in this paper. For instance, in [23], while ON/OFF Markov fluid arrivals are considered, wireless transmissions occur at fixed rates and wireless link is also modeled as a continuous-time Markov chain with ON and OFF states. In [30], a Gaussian autoregressive source is considered in the analysis. Additionally, our analysis, maximum average arrival rate expressions, and throughput characterizations in the low- θ , high- and low-SNR regimes are novel contributions with respect to these prior studies.

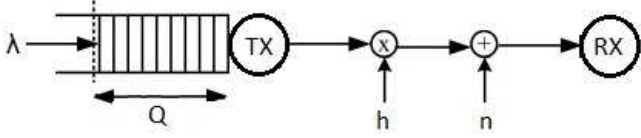


Fig. 1. System Model.

can be expressed as

$$y_i = h_i x_i + n_i \text{ for } i = 1, 2, \dots, m \quad (1)$$

where x_i and y_i are the channel input and output, respectively. The average energy of the input is \mathcal{E} , i.e.,

$$\mathbb{E}\{|x_i|^2\} = \mathcal{E}. \quad (2)$$

n_i denotes the zero-mean, circularly-symmetric, complex Gaussian noise with variance $\mathbb{E}\{|n_i|^2\} = N_0$. Hence, the signal-to-noise ratio is

$$\text{SNR} = \frac{\mathbb{E}\{|x|^2\}}{\mathbb{E}\{|n|^2\}} = \frac{\mathcal{E}}{N_0}. \quad (3)$$

Above in (1), h_i denotes the fading coefficient. Fading coefficients are assumed to be identically distributed, and the fading distribution can be arbitrary with finite variance.

While the ensuing analysis is applicable to a general class of fading distributions, we use a Gauss-Markov fading model in the numerical results and assume that the Gaussian fading coefficients in each block of m symbols follow the correlation pattern $h_i = \rho h_{i-1} + w_i$ where w_i is an independent, zero-mean Gaussian random variable with variance $\mathbb{E}\{|w_i|^2\} = (1 - \rho^2)\sigma_h^2$, $\rho \in [0, 1]$, and σ_h^2 is the common variance of the fading coefficients $\{h_i\}$. Note that when $\rho = 1$, we have full correlation, whereas $\rho = 0$ models the case of independent fading.

B. Queueing Constraints

We assume that the data to be transmitted is generated from random sources and is first stored in a buffer before transmission. Statistical constraints are imposed on the queue length. In particular, we assume that the buffer violation/overflow probability satisfies

$$\lim_{q \rightarrow \infty} \frac{\log \Pr\{Q \geq q\}}{q} = -\theta \quad (4)$$

where Q denotes the stationary queue length, and θ is the decay rate of the tail distribution of the queue length. The above limiting formula implies that for large q , we have

$$\Pr\{Q \geq q\} \approx e^{-\theta q}. \quad (5)$$

Indeed, a closer approximation is [20]

$$\Pr\{Q \geq q\} \approx \varsigma e^{-\theta q} \quad (6)$$

where $\varsigma = \Pr\{Q > 0\}$ is the probability of non-empty buffer². From (6), we notice that, for a sufficiently large threshold, the

buffer overflow probability should decay exponentially with rate controlled by the QoS exponent θ . Note that as θ increases, stricter queueing or QoS constraints are imposed, while looser queueing constraints are implied by smaller values of θ . Conversely, for a given buffer threshold q and overflow probability limit $\epsilon = \Pr\{Q \geq q\}$, the desired value of θ can be determined as

$$\theta = \frac{1}{q} \log_e \frac{\varsigma}{\epsilon}. \quad (7)$$

In the given setting, the delay violation probability is also characterized to decay exponentially and is approximated by [31]

$$\Pr\{D \geq d\} \approx \varsigma e^{-\theta a^*(\theta)d} \quad (8)$$

where D is the queueing delay in the buffer at steady state, d is the delay threshold, and $a^*(\theta)$ is the effective bandwidth of the arrival process, described below.

Next, we introduce the notions of effective bandwidth and effective capacity which we subsequently employ to formulate the wireless throughput in fading channels in the presence of random arrivals and statistical queueing constraints.

1) *Effective Bandwidth*: Effective bandwidth characterizes the minimum constant transmission (or service) rate required to support the given random data arrival process while the buffer overflow probability is limited or more explicitly the statistical queueing constraint described by (4) is satisfied. Let $\{a(k), k = 1, 2, \dots\}$ be a sequence of nonnegative random variables, describing the random arrival rates. Let also the time-accumulated arrival process be denoted by $A(t) = \sum_{k=1}^t a(k)$. Then, the effective bandwidth is given by the asymptotic logarithmic moment generating function of $A(t)$ [8], i.e.,

$$a^*(\theta) = \lim_{t \rightarrow \infty} \frac{1}{\theta t} \log \mathbb{E} \left\{ e^{\theta A(t)} \right\}. \quad (9)$$

In Section II-C, we describe the effective bandwidth of different source arrival models in detail.

2) *Effective Capacity*: Effective capacity, as a dual concept to effective bandwidth, identifies the maximum constant arrival rate that can be supported by a given time-varying service process while satisfying (4). Let $\{\nu[k], k = 1, 2, \dots\}$ denote the discrete-time stationary and ergodic stochastic service process and $S[t] \triangleq \sum_{k=1}^t \nu[k]$ be the time-accumulated service process. Then, the effective capacity is given by [20]

$$C_E(\text{SNR}, \theta) = - \lim_{t \rightarrow \infty} \frac{1}{\theta t} \log_e \mathbb{E} \left\{ e^{-\theta S[t]} \right\}. \quad (10)$$

Note that we have assumed that the fading coefficients $\{h_i\}$ change independently from one block of m symbols to another. Under this assumption, effective capacity simplifies to

$$C_E(\text{SNR}, \theta) = - \frac{1}{\theta} \log_e \mathbb{E} \left\{ e^{-\theta \nu} \right\} \quad (11)$$

where ν is the instantaneous service (or equivalently transmission) rate in one block. If the channel input sequence $\{x_i\}$ is an independent and identically distributed (i.i.d.) sequence of Gaussian random variables with zero mean and variance \mathcal{E} , then the service rate can be written as

$$\nu = \sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \quad (12)$$

²Probability of non-empty buffer can be approximated from the ratio of average arrival rate to average service rate [33].

where we have defined $z_i = |h_i|^2$. Hence, the effective capacity in the units of bits/block is

$$C_E(\text{SNR}, \theta) = -\frac{1}{\theta} \log_e \mathbb{E} \left\{ e^{-\theta \sum_{i=1}^m \log_2(1+\text{SNR}z_i)} \right\}. \quad (13)$$

Remark 1: In the special case of independent channel coefficients in each block and Rayleigh fading, we can express the effective capacity in closed-form as

$$C_E(\text{SNR}, \theta) = -\frac{m}{\theta} \log_e \left[\text{SNR}^{-\frac{\theta}{\log_e 2}} e^{\frac{1}{\text{SNR}}} \Gamma \left(1 - \frac{\theta}{\log_e 2}, \frac{1}{\text{SNR}} \right) \right] \quad (14)$$

$$= m \log_2(\text{SNR}) - \frac{m}{\theta \text{SNR}} - \frac{m}{\theta} \log_e \Gamma \left(1 - \frac{\theta}{\log_e 2}, \frac{1}{\text{SNR}} \right) \quad (15)$$

where $\Gamma(s, w) = \int_w^\infty \tau^{s-1} e^{-\tau} d\tau$ is the upper incomplete gamma function.

C. Effective Bandwidths of Different Source Models

1) *Discrete-Time Markov Sources:* In this subsection, we consider discrete-time Markov source models. Assume that the transition probability matrix of the n -state irreducible and aperiodic Markov source process is denoted by \mathbf{J} , and λ_i is the arrival rate in state i . Moreover, $\mathbf{\Lambda} = \text{diag} \{ \lambda_1, \lambda_2, \dots, \lambda_n \}$ is the diagonal matrix of arrival rates. Then, the effective bandwidth of this discrete Markov source is given by [11]

$$a(\theta) = \frac{1}{\theta} \log_e [\text{sp}(e^{\theta \mathbf{\Lambda}} \mathbf{J})] \quad (16)$$

where $\text{sp}(\cdot)$ is the spectral radius of the input matrix. Note that the stationary distribution $\boldsymbol{\pi}$ can be found from the solution of

$$\begin{aligned} \boldsymbol{\pi} \mathbf{1} &= 1, \\ \boldsymbol{\pi} \mathbf{J} &= \boldsymbol{\pi} \end{aligned} \quad (17)$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_n]$ and $\mathbf{1} = [1, \dots, 1]^T$.

In order to unveil the key relationships and tradeoffs, we consider a particularly simple two-state model. We assume that data arrival is either in the ON or OFF state in each block duration of m symbols. When the state is ON, λ bits arrive (i.e., the arrival rate is λ bits/block), while there are no arrivals in the OFF state. For this two-state model, the state transition probability matrix is given as

$$\mathbf{J} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}. \quad (18)$$

Given the above transition matrix \mathbf{J} , the effective bandwidth for this ON-OFF Markov model can be derived as [11]

$$a^*(\theta, \lambda) = \frac{1}{\theta} \log_e \left(\frac{p_{11} + p_{22} e^{\lambda \theta} + \sqrt{(p_{11} + p_{22} e^{\lambda \theta})^2 - 4(p_{11} + p_{22} - 1)e^{\lambda \theta}}}{2} \right) \quad (19)$$

where p_{11} denotes the probability of staying in the OFF state from one block to another. Similarly, p_{22} denotes the probability of staying in the ON state. The probabilities of transitioning from one state to a different one are therefore denoted by $p_{21} = 1 - p_{22}$

and $p_{12} = 1 - p_{11}$. For these transition probabilities, we can easily see that the probability of the ON state in the steady state is

$$P_{\text{ON}} = \frac{1 - p_{11}}{2 - p_{11} - p_{22}}. \quad (20)$$

Therefore, the average arrival rate is

$$r_{\text{avg}} = \lambda P_{\text{ON}} = \lambda \frac{1 - p_{11}}{2 - p_{11} - p_{22}} \quad (21)$$

which is equal to the average departure rate when the queue is in steady state [39].

2) *Markov Fluid Sources:* In this subsection, we address Markov fluid sources where the source arrival process is modeled as a continuous-time Markov chain. Assume that \mathbf{G} is the irreducible transition rate matrix of the Markov chain, λ_i is the arrival rate in the i^{th} state, and $\mathbf{\Lambda} = \text{diag} \{ \lambda_1, \lambda_2, \dots, \lambda_n \}$. Then, the effective bandwidth of this source is given by [14], [15]

$$a^*(\theta) = \mu \left(\mathbf{\Lambda} + \frac{1}{\theta} \mathbf{G} \right) \quad (22)$$

where $\mu(\cdot)$ denotes the maximum real eigenvalue of the input matrix. We also note that the stationary distribution $\boldsymbol{\pi}$ of the continuous-time Markov chain can be found by solving

$$\begin{aligned} \boldsymbol{\pi} \mathbf{1} &= 1, \\ \boldsymbol{\pi} \mathbf{G} &= \mathbf{0} \end{aligned} \quad (23)$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_n]$, $\mathbf{0} = [0, \dots, 0]^T$ and $\mathbf{1} = [1, \dots, 1]^T$.

In order to derive closed-form expressions in our analysis, we again consider two states (ON/OFF). When there is no arrival, the state is OFF. When the state is ON, the arrival rate is λ bits/block. The transition rate matrix for a two-state Markov fluid is in the form of

$$\mathbf{G} = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}, \quad (24)$$

where α is the transition rate from OFF state to ON state whereas β is the transition rate from ON state to OFF state. Using (22), we can express the effective bandwidth as

$$a^*(\theta) = \frac{1}{2\theta} \left[\theta \lambda - (\alpha + \beta) + \sqrt{(\theta \lambda - (\alpha + \beta))^2 + 4\alpha\theta\lambda} \right]. \quad (25)$$

The probability of ON state, π_2 , is required to define the average rate. Inserting the generator matrix \mathbf{G} in (24) into (23), we obtain the ON state probability as

$$\pi_2 = P_{\text{ON}} = \frac{\alpha}{\alpha + \beta}. \quad (26)$$

Therefore, the average arrival rate of the two-state Markov fluid process is

$$r_{\text{avg}} = \lambda P_{\text{ON}} = \lambda \frac{\alpha}{\alpha + \beta}. \quad (27)$$

3) *Markov Modulated Poisson Sources*: In this subsection, we assume that the data arrival to the buffer is a Poisson process whose intensity is controlled by a continuous-time Markov chain. For instance, the intensity of the Poisson arrival process is λ_i in the i^{th} state of the Markov chain. Therefore, the source arrival is modeled as a Markov-modulated Poisson process (MMPP). Assuming that the \mathbf{G} is the irreducible transition rate matrix of the Markov chain and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ is the diagonal matrix of the intensities of the Poisson arrivals in different states, the effective bandwidth is given by [14], [15]

$$a^*(\theta) = \frac{1}{\theta} \mu \left((e^\theta - 1) \mathbf{\Lambda} + \mathbf{G} \right). \quad (28)$$

As in previous sections, we consider a two-state (ON/OFF) model in which there are no arrivals in the OFF state (i.e., the intensity is 0) and the intensity of the Poisson arrival process is λ bits/block in the ON state. Assuming the same generator matrix \mathbf{G} as in (24), we can express the effective bandwidth as

$$a^*(\theta) = \frac{1}{\theta} \left[(e^\theta - 1) \lambda - (\alpha + \beta) \right] + \frac{1}{2\theta} \sqrt{\left[(e^\theta - 1) \lambda - (\alpha + \beta) \right]^2 + 4\alpha(e^\theta - 1)\lambda}. \quad (29)$$

Note that the average arrival rate in bits/block is again given by

$$r_{\text{avg}} = \lambda P_{\text{ON}} = \lambda \frac{\alpha}{\alpha + \beta}. \quad (30)$$

We further note that if the transition rate $\beta = 0$, then we have $P_{\text{ON}} = 1$. In this case, MMPP model specializes to a pure Poisson source with intensity λ bits/block, and the effective bandwidth of this source is given by

$$a^*(\theta) = \frac{1}{\theta} (e^\theta - 1) \lambda. \quad (31)$$

III. THROUGHPUT WITH MARKOVIAN SOURCE MODELS

In this section, we formulate the throughput of wireless fading channels when the data arrivals are random and statistical queueing constraints are imposed. More specifically, we consider Markovian arrival models introduced in Section II-C, namely discrete-time Markov sources, Markov fluids and Markov-modulated Poisson arrivals. The states in these Markov processes are differentiated by the corresponding arrival rates in these states, e.g., the arrival rate in the i^{th} state is λ_i . If the stationary distribution of the Markov process is denoted by π , the average arrival rate in an n -state Markov source model simply becomes

$$r_{\text{avg}} = \sum_{i=1}^n \pi_i \lambda_i \quad (32)$$

which is equal to the average departure rate when the queue is in steady state [39].

We seek to determine the throughput by identifying the maximum average arrival rate that can be supported by the fading channel described in Section II-A while satisfying the statistical QoS limitations given in the form in (4). As shown in [39, Theorem 2.1], (4) is satisfied, i.e., buffer violation probability decays exponentially fast with rate controlled by the QoS exponent θ ,

if the effective bandwidth of the arrival process is equal to the effective capacity of the service process, i.e.,

$$a^*(\theta) = C_E(\text{SNR}, \theta). \quad (33)$$

Hence by solving (33), we can determine the maximum average arrival rate $r_{\text{avg}}^*(\text{SNR}, \theta)$. By specifying the effective bandwidth of different source models and incorporating the effective capacity of time-varying wireless transmissions in (13), the maximum average arrival rate can be determined for general n -state Markovian source models. Indeed, several n -state source models are addressed in Section IV. However, in our analysis in this section, to illustrate the impact of the arrival and system parameters in a lucid setting, we concentrate on the two-state (ON-OFF) arrival models and provide closed-form expressions for the maximum average arrival rates in terms of the source parameters and the effective capacity of the wireless transmissions. We also identify the characteristics of the throughput in the low- θ and high-SNR regimes. We note that the analysis throughout this section is applicable to any arbitrary fading correlation within each fading block, with the exception of high-SNR characterizations which are obtained under the assumption of i.i.d. fading.

A. Discrete-Time Markov Sources

In this section, we consider two-state (ON/OFF) discrete Markov sources described in Section II-C1, and initially characterize the maximum average arrival rate r_{avg}^* that can be supported by the fading channel while satisfying the statistical QoS limitations given in the form in (4).

Theorem 1: For the two-state (ON/OFF) discrete Markov source, the maximum average arrival rate (in bits/block) as a function of the QoS exponent θ , effective capacity of the fading channel $C_E(\text{SNR}, \theta)$, and the state transition probabilities is expressed as

$$r_{\text{avg}}^*(\text{SNR}, \theta) = \frac{P_{\text{ON}}}{\theta} \log_e \left(\frac{e^{2\theta C_E(\text{SNR}, \theta)} - p_{11} e^{\theta C_E(\text{SNR}, \theta)}}{1 - p_{11} - p_{22} + p_{22} e^{\theta C_E(\text{SNR}, \theta)}} \right). \quad (34)$$

Proof: See Appendix A.

Note that r_{avg}^* above is formulated in terms of the effective capacity, C_E , of wireless transmissions. In Fig. 2, we plot the maximum average arrival rate as a function of the effective capacity for different source characteristics when $\theta = 1$. It is easy to verify that when $P_{\text{ON}} = \frac{1-p_{11}}{2-p_{11}-p_{22}} = 1$ or equivalently $p_{22} = 1$, (34) simplifies to $r_{\text{avg}}^*(\text{SNR}, \theta) = C_E(\text{SNR}, \theta)$. Hence, when the source is always ON and therefore the arrivals are at a constant rate, maximum average arrival rate is equal to the effective capacity, as also observed in Fig. 2. On the other hand, we notice in this figure that as P_{ON} diminishes and the source becomes more bursty, throughput diminishes as well and smaller average arrival rates are supported for given effective capacity.

As also indicated in the above discussion and seen in (34), $r_{\text{avg}}^*(\text{SNR}, \theta)$ is in general a function of the state transition probabilities of the Markov arrival process in the presence of buffer constraints. On the other hand, as shown in the following result, this dependence disappears if no buffer constraints are imposed, i.e., when $\theta = 0$.

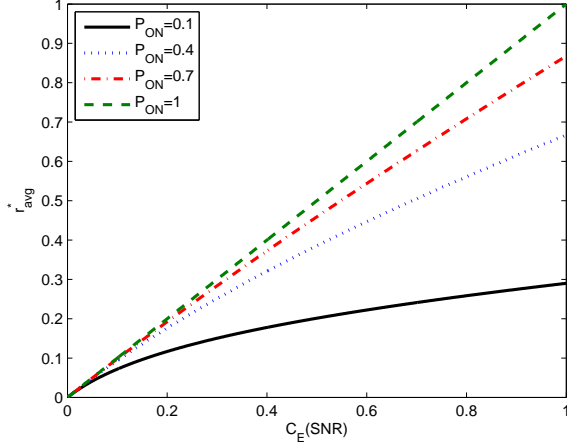


Fig. 2. Maximum average arrival rate r_{avg}^* vs. effective capacity $C_E(\text{SNR})$ for different source statistics. No fading correlation, i.e., $\rho = 0$. $\theta = 1$.

Theorem 2: As the statistical queueing constraints are relaxed by letting the QoS exponent θ approach zero, the maximum average arrival rate converges to

$$\lim_{\theta \rightarrow 0} r_{\text{avg}}^*(\text{SNR}, \theta) = \sum_{i=1}^m \mathbb{E} \{ \log_2(1 + \text{SNR} z_i) \} \text{ bits/block.} \quad (35)$$

Moreover, the first derivative of r_{avg}^* with respect to θ at $\theta = 0$ is

$$\begin{aligned} \left. \frac{\partial r_{\text{avg}}^*(\text{SNR}, \theta)}{\partial \theta} \right|_{\theta=0} &= -\frac{1}{2} \sum_{i,j=1}^m \text{cov} \{ \log_2(1 + \text{SNR} z_i), \log_2(1 + \text{SNR} z_j) \} \\ &\quad - \frac{\eta}{2} \left(\sum_{i=1}^m \mathbb{E} \{ \log_2(1 + \text{SNR} z_i) \} \right)^2 \end{aligned} \quad (36)$$

where we define η as

$$\eta = \frac{(1 - p_{22})(p_{11} + p_{22})}{(1 - p_{11})(2 - p_{11} - p_{22})}. \quad (37)$$

Proof: See Appendix B.

We see from (35) that if no statistical buffer constraints are imposed i.e., if $\theta = 0$, then the maximum average arrival rate is equal to the ergodic capacity of the block-fading channel, and therefore is independent of the statistical characteristics of the discrete Markov arrival model. Moreover, the dependence of the maximum arrival rate in this regime on the channel statistics is only through the marginal distributions of the fading coefficients. Hence, channel correlation in each fading block does not play any role. However, this radically changes when $\theta > 0$. For instance, we notice from (36) that even with a small increase in θ , r_{avg}^* starts varying with the source and channel statistics, as exemplified by the dependence of the first derivative on η and the covariance function.

Having discussed the low- θ regime above, we next provide a characterization of $r_{\text{avg}}^*(\text{SNR}, \theta)$ at high SNR values for i.i.d. Rayleigh fading.

Theorem 3: Assume that the channel fading coefficients are i.i.d. in each block and fading power $z = |h|^2$ is exponentially

distributed with unit mean (i.e., Rayleigh fading is experienced). Then, we have

$$\frac{1}{m} r_{\text{avg}}^*(\text{SNR}, \theta) = \begin{cases} \frac{P_{\text{ON}}}{\theta \log_2 e} \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } \theta > \frac{1}{\log_2 e} \\ P_{\text{ON}} \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } 0 < \theta < \frac{1}{\log_2 e} \\ \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } \theta = 0 \end{cases} \quad (38)$$

as $\text{SNR} \rightarrow \infty$.

Proof: See Appendix C.

Note that the high-SNR slope is defined as [37]

$$\mathcal{S}_{\infty} = \lim_{\text{SNR} \rightarrow \infty} \frac{\frac{1}{m} r_{\text{avg}}^*(\text{SNR}, \theta)}{\log_2 \text{SNR}}. \quad (39)$$

Theorem 3 shows that the high-SNR slope of the maximum arrival rate for the two-state discrete Markov source that can be supported in the i.i.d Rayleigh fading channel is

$$\mathcal{S}_{\infty} = \begin{cases} \frac{P_{\text{ON}}}{\theta \log_2 e} & \text{if } \theta > \frac{1}{\log_2 e} \\ P_{\text{ON}} & \text{if } 0 < \theta < \frac{1}{\log_2 e} \\ 1 & \text{if } \theta = 0 \end{cases}. \quad (40)$$

It is interesting to observe from Theorem 2 that when no buffer constraints are imposed i.e., when $\theta = 0$, the high-SNR slope is $\mathcal{S}_{\infty} = 1$, again independent of the source statistics. On the other hand, when $\theta > 0$, \mathcal{S}_{∞} becomes proportional to the ON probability and is now less than one unless the arrival rate is constant. Furthermore, for θ values greater than $\frac{1}{\log_2 e}$, \mathcal{S}_{∞} starts decreasing with increasing θ . Hence, the result in Theorem 3 quantifies the performance degradation experienced at high SNR levels due to source randomness and statistical buffer constraints.

Let us further simplify the source model and set $p_{11} = 1 - s$ and $p_{22} = s$. The source is now described by the single parameter s . Notice that with this choice we have $P_{\text{ON}} = s$ and hence s becomes a measure of the burstiness of the source. The smaller the s , the less frequently the data arrives and the more bursty the source becomes. At the other extreme, if $s = 1$, source is ON all the time and we have constant arrival rate. Furthermore, with the above choice of p_{11} and p_{22} , the expression for the maximum average arrival rate simplifies to

$$r_{\text{avg}}^*(\text{SNR}, \theta) = \frac{s}{\theta} \log_e \left(\frac{e^{\theta C_E(\text{SNR}, \theta)} - (1 - s)}{s} \right), \quad (41)$$

which can readily be seen to be a diminishing function as s decreases. Therefore, source burstiness generally hurts the throughput if we keep all other variables fixed.

We can further observe this in Fig. 3, where we plot the maximum average arrival rate (or equivalently the throughput) as a function of SNR for different values of s and the QoS exponent θ . Numerical analysis verifies that as the source becomes more bursty with lower values of s , throughput diminishes. Conversely, throughput is maximized when $s = 1$ i.e., when we have constant arrival rates. It is also interesting to notice from (41) that the arrival rate in the ON state, which is given by $\lambda^* = \frac{r_{\text{avg}}^*(\text{SNR}, \theta)}{s}$, increases as s diminishes. Hence, smaller s implies that data arrives less frequently but with bursts of increased rates. We also observe in Fig. 3 that the throughput reduction due to burstiness is more severe at high SNRs. This is indeed a consequence of the fact that high-SNR slope gets smaller as $P_{\text{ON}} = s$ decreases,

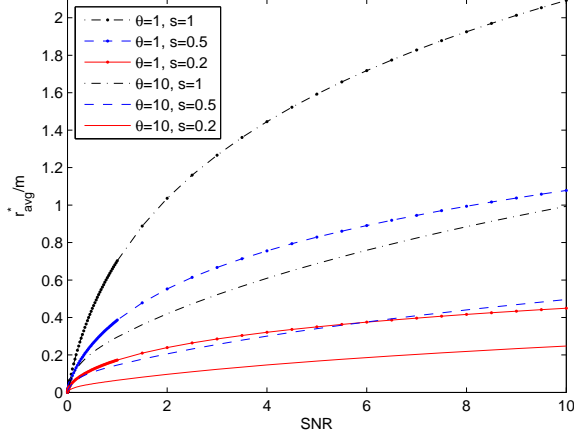


Fig. 3. Maximum average arrival rate r_{avg}^* vs. signal-to-noise ratio SNR for different values of θ and source statistics. No fading correlation, i.e., $\rho = 0$.

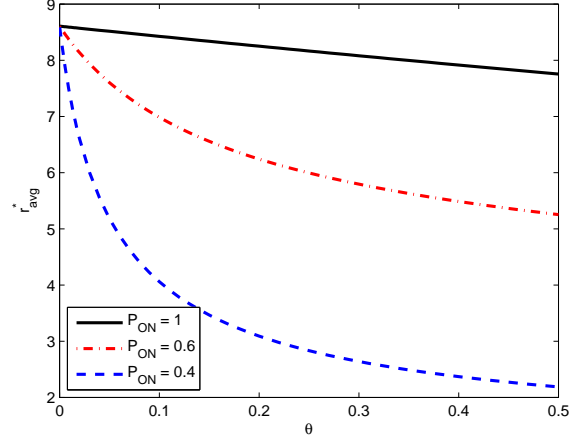


Fig. 5. Maximum average arrival rate r_{avg}^* vs. QoS exponent θ for different values of P_{ON} . No fading correlation, i.e., $\rho = 0$.

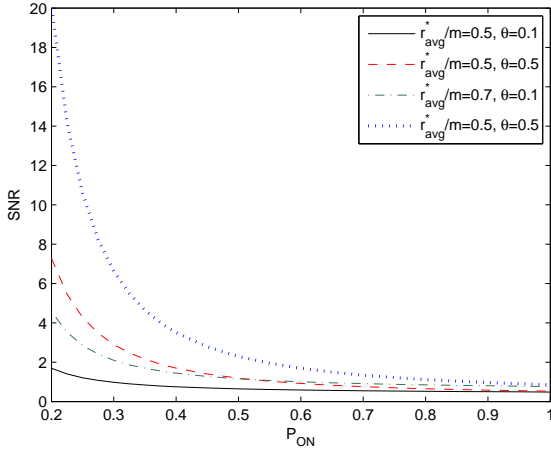


Fig. 4. Required SNR vs. ON probability, P_{ON} , for a given fixed average arrival rate. No fading correlation, i.e., $\rho = 0$.

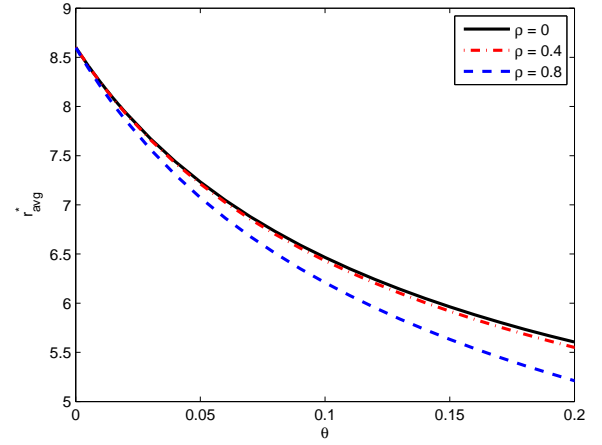


Fig. 6. Maximum average arrival rate r_{avg}^* vs. QoS exponent θ for different fading correlations. $P_{\text{ON}} = 0.5$.

as discussed above. Finally, we see in Fig. 3 that performance degradation is experienced as θ increases and hence stricter buffer constraints are imposed.

In Fig. 4, we plot the SNR levels required to support a given average arrival rate as a function of the ON-state probability for different values of the QoS exponent θ . We observe that as P_{ON} decreases and hence the source becomes more bursty, required SNR level increases in general. Interestingly, a sharper increase is experienced under stricter buffer constraints (e.g., when $\theta = 0.5$ rather than $\theta = 0.1$), indicating higher power/energy costs in these cases.

The low- θ regime is investigated in Fig. 5 where we plot the maximum average arrival rate r_{avg}^* vs. QoS exponent θ for different P_{ON} values. We set $\text{SNR} = 1$. We notice that all three curves converge to the same throughput value $r_{\text{avg}}^*(0)$ as $\theta \rightarrow 0$, confirming the result in (35). Hence, source characteristics do not affect the throughput if no queuing constraints are imposed. As θ increases, throughput diminishes and the reduction in r_{avg}^* is

more severe for more bursty sources (e.g., when $P_{\text{ON}} = 0.4$). We notice that, as predicted by (36), this is already reflected by the different slopes of r_{avg}^* in the vicinity of $\theta = 0$. Hence, overall the system for more bursty sources becomes more cautious and supports smaller average arrival rates in order to avoid buffer overflows.

In Fig. 6, we again plot the throughput as a function θ but for different values of ρ , which quantifies the correlation between fading coefficients in each fading block. We fix $\text{SNR} = 1$ and set $P_{\text{ON}} = 0.5$. Similar to burstiness, fading correlation does not have any effect on the throughput when $\theta = 0$. When $\theta > 0$, higher correlation (i.e., larger ρ) results in lower supported throughput under the same QoS constraints.

Finally, we have conducted simulations to further verify the theoretical analysis and results. In particular, in the simulations, for fixed QoS exponent θ , SNR, and state transition probabilities p_{11} and p_{22} of the ON/OFF discrete Markov source, we initially determine the maximum average arrival rate from (34) and the

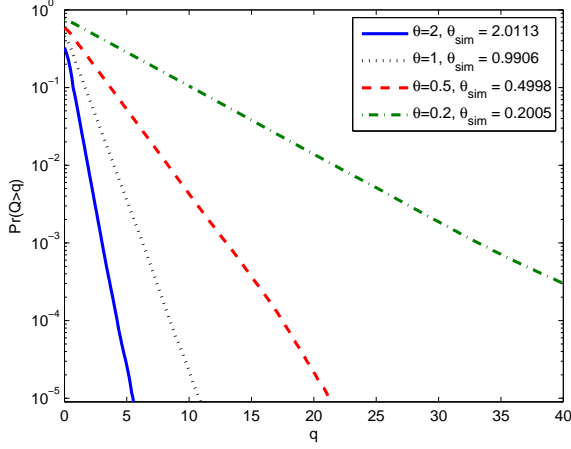


Fig. 7. Buffer overflow probability $\Pr\{Q > q\}$ vs. buffer threshold q for different values of θ . $p_{11} = p_{22} = 0.8$, SNR = 0 dB

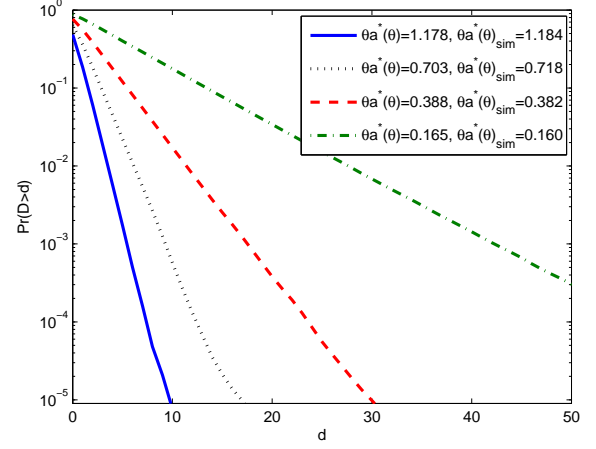


Fig. 8. Delay violation probability $\Pr\{D > d\}$ vs. delay threshold d for different values of θ . $p_{11} = p_{22} = 0.8$, SNR = 0 dB

corresponding maximum arrival rate in the ON state. Then, using the given statistical characterizations and the maximum arrival rate, we generate random Markov arrivals and assume that the arriving data is initially stored in the buffer before being transmitted. Transmission rates are simulated by generating realizations of i.i.d. Gaussian fading coefficients. Throughout this process, we track the queue evolution and the buffer state (i.e., the queue length) as the Markov arrivals occur (and hence more data gets stored) and transmissions at varying rates according to the generated fading coefficients are performed, clearing some data off the buffer. In Figs. 7 and 8, we plot the simulated buffer overflow probability $\Pr\{Q \geq q\}$ and delay violation probability $\Pr\{D > d\}$, respectively, as functions of the corresponding thresholds, following 10^7 runs of the simulation. We notice that while the theoretical analysis makes use of results from the theory of large deviations and is generally applicable for large thresholds, the simulation results are interestingly in excellent agreement with the theoretical predictions even at small values of the thresholds. For instance, we note from (6) that $\log_e \Pr\{Q \geq q\} \approx \log_e \varsigma - \theta q$ and hence is expected to decay linearly in q with slope θ . We indeed observe this linear decay in Fig. 7 (where the overflow probabilities are plotted in logarithmic scale) for even small to moderate values of q . Moreover, the slopes of the simulated curves, denoted by θ_{sim} , are very close to the originally selected value of θ . Similar conclusions apply to Fig. 8 as well. In this figure, delay violation probabilities are determined by keeping track of the delay experienced by the data stored in the buffer until transmission. We again notice that the logarithm of the delay violation probability decays linearly with threshold d (or equivalently the delay violation probability diminishes exponentially with d). Note that the slope of the linear decay is predicted from (8) to be $\theta a^*(\theta)$ where $a^*(\theta)$ is the effective bandwidth of the source. Again, the slope of the simulated curves are almost the same as this theoretical slope value, as indicated in the legend on the figure.

B. Markov Fluid Sources

In this section, we consider Markov fluid sources. In the following, we go through similar steps as in the previous subsection and initially determine the maximum average arrival rates of ON/OFF Markov fluid sources that can be supported by the wireless channel as a function of the source transition rates and the effective capacity of wireless transmissions. Subsequently, we give characterizations of the maximum average arrival rates in the low- θ and high-SNR regimes.

Theorem 4: For the two-state (ON/OFF) Markov fluid source, the maximum average arrival rate is given as

$$r_{\text{avg}}^*(\text{SNR}, \theta) = P_{\text{ON}} \frac{\theta C_E(\text{SNR}, \theta) + \alpha + \beta}{\theta C_E(\text{SNR}, \theta) + \alpha} C_E(\text{SNR}, \theta). \quad (42)$$

Proof: See Appendix D.

Note that maximum average arrival rate generally depends on the transition rate matrix of the Markov fluid source. At the same time, similar to the discrete case, when there are no QoS constraints, source characteristics do not have any impact on the throughput. However, this changes drastically when $\theta > 0$ even if θ is vanishingly small. These properties are demonstrated analytically in the result below.

Theorem 5: As the statistical queueing constraints are relaxed by letting the QoS exponent θ approach zero, we have

$$\lim_{\theta \rightarrow 0} r_{\text{avg}}^*(\text{SNR}, \theta) = \sum_{i=1}^m \mathbb{E} \{ \log_2(1 + \text{SNR} z_i) \} \text{ bits/block}, \quad (43)$$

and

$$\left. \frac{\partial r_{\text{avg}}^*(\text{SNR}, \theta)}{\partial \theta} \right|_{\theta=0} = -\frac{1}{2} \sum_{i,j=1}^m \text{cov} \{ \log_2(1 + \text{SNR} z_i), \log_2(1 + \text{SNR} z_j) \} - \frac{\zeta}{2} \left(\sum_{i=1}^m \mathbb{E} \{ \log_2(1 + \text{SNR} z_i) \} \right)^2 \quad (44)$$

where ζ is defined as

$$\zeta = \frac{2\beta}{\alpha(\alpha + \beta)}. \quad (45)$$

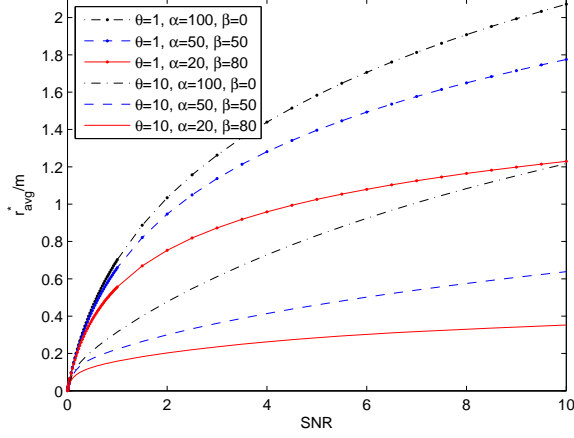


Fig. 9. Maximum average arrival rate r_{avg}^* vs. signal-to-noise ratio SNR for different values of θ and source statistics. No fading correlation, i.e., $\rho = 0$.

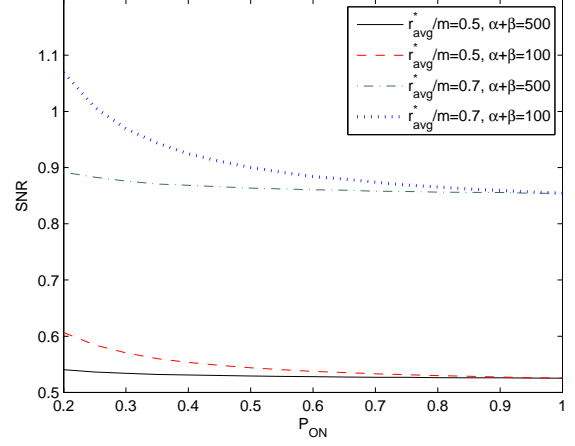


Fig. 10. Required SNR vs. ON probability, P_{ON} , for a given average arrival rate. $\theta = 0.5$. No fading correlation, i.e., $\rho = 0$.

Proof: See Appendix E.

We note that when $\theta > 0$, r_{avg}^* depends on the source and channel statistics. In (44), we observe the dependence of even the first derivative on channel correlations and source statistics via the covariance function and the parameter ζ , respectively.

Next we present a high-SNR characterization of the throughput for Rayleigh fading.

Theorem 6: Assume that the channel fading coefficients are i.i.d. in each block and fading power $z = |h|^2$ is exponentially distributed with unit mean (i.e., Rayleigh fading is experienced). Then, we have

$$\frac{1}{m} r_{\text{avg}}^*(\text{SNR}, \theta) = \begin{cases} \frac{P_{\text{ON}}}{\theta \log_2 e} \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } \theta > \frac{1}{\log_2 e} \\ P_{\text{ON}} \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } 0 < \theta < \frac{1}{\log_2 e} \\ \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } \theta = 0 \end{cases} \quad (46)$$

as $\text{SNR} \rightarrow \infty$.

The proof of Theorem 6 is omitted due to its similarity to the proof of Theorem 3 in Appendix C. Similar conclusions as in Section II-C1 immediately apply.

Note that the throughput expression in (42) suggests that for sufficiently high SNR levels leading to $\theta C_E(\text{SNR}, \theta) \gg \alpha + \beta$, we have

$$r_{\text{avg}}^*(\text{SNR}, \theta) \approx P_{\text{ON}} C_E(\text{SNR}, \theta). \quad (47)$$

Hence, at high SNRs, the maximum average arrival rate depends on the source statistics only through the ON probability. This is noted in the high-SNR behavior in (46) as well.

In Fig. 9, we plot r_{avg}^* vs. SNR curves for different α , β , and θ values. We immediately observe that throughput diminishes with increasing θ and decreasing P_{ON} . In Fig. 10, we analyze the effect of r_{avg} , P_{ON} , and $\alpha + \beta$ on the required SNR levels. For Markov fluid sources, ON state probability is not the sole indicator of burstiness. Having low α and β values also indicates that source is more bursty as the transition between ON and OFF states becomes less frequent. Hence, OFF state can be more persistent. When α and β are large, state transitions occur more

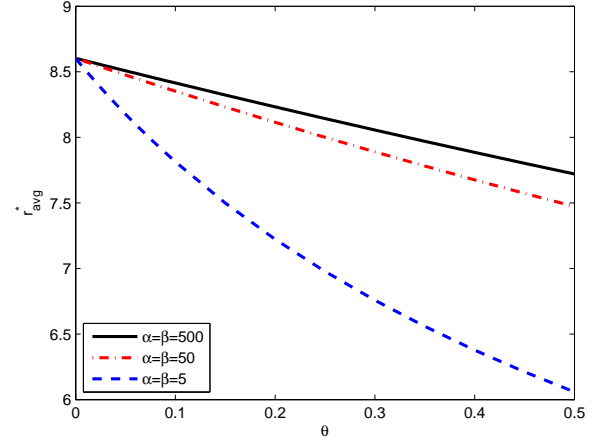


Fig. 11. Maximum average arrival rate r_{avg}^* vs. QoS exponent θ for different values of α and β . $P_{\text{ON}} = 0.5$, $\rho = 0$, and $\text{SNR} = 0$ dB.

rapidly, leading to lower required SNR levels. Again, we notice that the burstiness is harmful for the system.

In Fig. 11, we plot the maximum average arrival rate r_{avg}^* as a function of θ for different values of α and β . Notice that by keeping $\alpha = \beta$, the ON probability P_{ON} is fixed at 0.5, while average durations of ON and OFF states vary as the values of $\alpha = \beta$ change. For example, higher α and β values lead to shorter periods for ON and OFF states on average. As an outcome of this fact, we observe in the figure that higher throughput is achieved with sources having higher $\alpha + \beta$.

C. Markov Modulated Poisson Sources

In this section, we address two-state (ON/OFF) MMPP sources. Similarly as for the previous source models, we determine the maximum average arrival rate of the MMPP source, which can be supported by the fading channel in the presence of QoS constraints, and investigate the throughput in the low- θ and

high-SNR regimes. The results can be immediately specialized to pure Poisson sources by setting $\beta = 0$.

Theorem 7: For the two-state (ON/OFF) MMPP source model, the maximum average arrival rate is

$$r_{\text{avg}}^*(\text{SNR}, \theta) = P_{\text{ON}} \frac{\theta [\theta C_E(\text{SNR}, \theta) + \alpha + \beta]}{(e^\theta - 1) [\theta C_E(\text{SNR}, \theta) + \alpha]} C_E(\text{SNR}, \theta). \quad (48)$$

Proof: See Appendix F.

It is interesting to observe that the throughput with the MMPP source is almost identical to that with the Markov fluid source model, save only for the multiplicative factor $\frac{\theta}{e^\theta - 1}$ in (48). Note that $\frac{\theta}{e^\theta - 1} < 1$ for $\theta > 0$ and diminishes exponentially fast with increasing θ . Hence, the throughput is generally smaller with MMPP sources and decreases fast with θ . This can be attributed to the much more randomness/burstiness we experience with an MMPP source with respect to the previous Markov models. Note that the arrival rate in the ON state, rather than being a constant as in the previous cases, is determined by a Poisson process. Hence, the presence of the term $\frac{\theta}{e^\theta - 1}$ is due to this Poisson property. Indeed, if we have a pure Poisson source, the maximum average arrival rate is $r_{\text{avg}}^*(\text{SNR}, \theta) = \frac{\theta}{(e^\theta - 1)} C_E(\text{SNR}, \theta)$ obtained by setting $\beta = 0$. The cost of this additional randomness is reflected in the following results as well.

Theorem 8: As the statistical queueing constraints are relaxed by letting the QoS exponent θ approach zero, we have

$$\lim_{\theta \rightarrow 0} r_{\text{avg}}^*(\text{SNR}, \theta) = \sum_{i=1}^m \mathbb{E} \{ \log_2(1 + \text{SNR} z_i) \} \text{ bits/block}, \quad (49)$$

and

$$\begin{aligned} \left. \frac{\partial r_{\text{avg}}^*(\text{SNR}, \theta)}{\partial \theta} \right|_{\theta=0} &= -\frac{1}{2} \sum_{i,j=1}^m \text{cov} \{ \log_2(1 + \text{SNR} z_i), \log_2(1 + \text{SNR} z_j) \} \\ &\quad - \frac{\zeta}{2} \left(\sum_{i=1}^m \mathbb{E} \{ \log_2(1 + \text{SNR} z_i) \} \right)^2 \\ &\quad - \frac{1}{2} \sum_{i=1}^m \mathbb{E} \{ \log_2(1 + \text{SNR} z_i) \} \end{aligned} \quad (50)$$

where

$$\zeta = \frac{2\beta}{\alpha(\alpha + \beta)}. \quad (51)$$

Proof: See Appendix G.

When the system is free of QoS limitations, the maximum average arrival rate for the MMPP source again turns out to be equal to the ergodic capacity. However, the throughput has a steeper decline in the low- θ regime due to the third term on the right-hand side of (50).

Theorem 9: Assume that the channel fading coefficients are i.i.d. in each block and fading power $z = |h|^2$ is exponentially distributed with unit mean (i.e., Rayleigh fading is experienced). Then, we have

$$\frac{1}{m} r_{\text{avg}}^*(\text{SNR}, \theta) = \begin{cases} \frac{P_{\text{ON}}}{(e^\theta - 1) \log_2 e} \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } \theta > \frac{1}{\log_2 e} \\ \frac{\theta}{e^\theta - 1} P_{\text{ON}} \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } 0 < \theta < \frac{1}{\log_2 e} \\ \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } \theta = 0 \end{cases} \quad (52)$$

as $\text{SNR} \rightarrow \infty$.

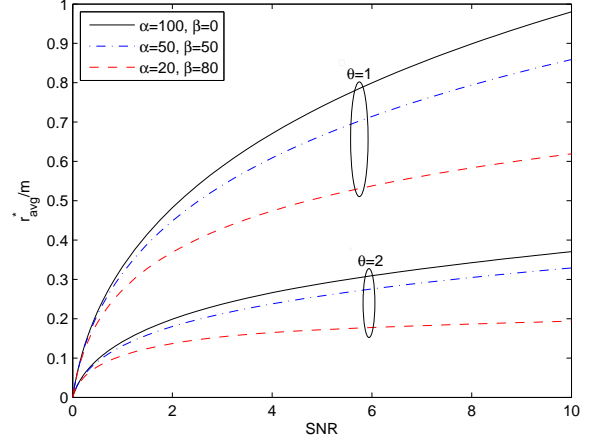


Fig. 12. Maximum average arrival rate r_{avg}^* vs. signal-to-noise ratio SNR for different values of θ and different source statistics. No fading correlation, i.e., $\rho = 0$.

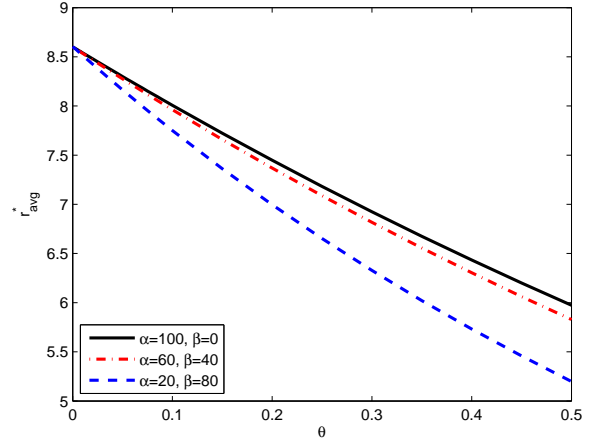


Fig. 13. Maximum average arrival rate r_{avg}^* vs. QoS exponent θ for different source statistics. $\rho = 0$.

Since the ratio between the MMPP throughput and Markov fluid throughput always stays at $\frac{\theta}{e^\theta - 1}$, we can immediately obtain the above high-SNR characterization, using the formulations in (46).

In the numerical results, we have similar conclusions as in the Markov fluid case. The primary difference is the reduced throughput for given θ , which, for instance, is readily seen when we compare Figs. 9 and 12, where we have throughput vs. SNR curves for Markov fluid and MMPP sources, respectively. In Fig. 13, we display the maximum average arrival rate r_{avg}^* as a function of θ . We set $\alpha + \beta = 100$ and $\text{SNR} = 1$, and vary α and β and hence the ON probability. We note that as P_{ON} decreases, the performance degrades faster with increasing θ , as indicated by the steeper slopes.

D. Comparative View of Source Models and Performance Levels

In our analysis, we have considered discrete-time Markov, Markov fluid, and MMPP arrival models. All models possess the Markovian property in the sense that the evolution of the Markov chains and hence the state transitions satisfy the Markov condition and are described by the transition probability matrix in the case of discrete-time models and by the transition rate matrix in the case of fluid (or equivalently continuous-time) models. Also, state holding times are geometrically distributed in discrete-time models and exponentially distributed in continuous-time models, and hence exhibit the memoryless property.

At the same time, there are distinct differences between different source models. For instance, transitions between states occur in discrete time steps in discrete-time Markov models while the Markov chain can spend a continuous amount of time in any state in Markov fluid models (i.e., the length of time spent in any state is a continuous random variable or more explicitly holding times are exponentially distributed as also noted above). MMPP models are further differentiated. In the discrete-time Markov and Markov fluid models, arrival rates are assumed to be constant in any given state. On the other hand, when the arrivals are modeled as MMPP, arrival rate is Poisson distributed in each state with a different intensity. Hence, MMPP sources exhibit a higher level of variation in this sense and can be regarded as a more bursty source.

We also remark that ON/OFF discrete-time Markov and Markov fluid source models can be easily specialized to the source with a constant arrival rate by letting ON state probability $P_{\text{ON}} = 1$. On the other hand, when $P_{\text{ON}} = 1$ in the ON/OFF MMPP source, we have a pure Poisson arrival source.

Finally, we note that although there is a certain degree of similarity in the analysis of discrete-time Markov and Markov fluid models and their throughput performances (e.g., high-SNR characterizations are the same in Theorems 3 and 6), the set of results for one model do not immediately follow from those for the other model as seen in the throughput formulations in (34) and (42) and the definitions of η and ζ in (37) and (45), respectively. However, there is a clear distinction when MMPP sources are considered. As also discussed in Section III-C, higher level of burstiness of MMPP sources penalizes the performance, and lower throughput levels are achieved in general with these sources.

IV. ENERGY EFFICIENCY ANALYSIS

In this section, we conduct a low-SNR analysis and investigate the energy efficiency in fading channels when data arrivals are random and statistical queueing constraints are imposed. We first identify the energy efficiency metrics. Subsequently, we consider different source arrival models and provide closed-form expressions for the energy efficiency metrics when the arrival rate is constant or follows a two-state Markovian model. We also numerically analyze specific n -state Markovian sources. Similarly as in the previous section, arbitrary fading correlation within each fading block is considered in the analysis.

A. Energy Efficiency Metrics

Before defining the energy efficiency metrics, we briefly describe the concavity of the maximum average arrival rate as

a function of SNR in the two-state (ON/OFF) arrival models (or if the arrival rates in an n -state model can be expressed as multiples of a certain single rate). In [32, Lemma 1], it was proven that effective capacity is a concave function of SNR. Elwalid and Mitra [14] showed that the effective bandwidth of a source is monotonically increasing when any arrival rate λ_i increases and is convex in the arrival rates $\{\lambda_1, \lambda_2, \dots, \lambda_N\}$. In the ON/OFF arrival models, we have a single arrival rate λ . Since effective bandwidth is a monotonically increasing and convex function of λ , the inverse function of the effective bandwidth a^{*-1} exists and is a nondecreasing concave function. More specifically, the maximum arrival rate can be expressed as $\lambda^*(\text{SNR}, \theta) = a^{*-1}(C_E(\text{SNR}, \theta))$, which is a nondecreasing concave function of the effective capacity, which is concave in SNR. Using the composition properties of concave functions [40], we realize that the maximum arrival rate is concave in SNR. Thus, the maximum average arrival rate $r_{\text{avg}}^*(\text{SNR}, \theta)$ is also concave in SNR.

In our analysis, following the approach in [41], we study the minimum energy per bit and the wideband slope, which is defined as the slope of the spectral efficiency curve at zero spectral efficiency, as the performance metrics of energy efficiency. While minimum bit energy is a performance measure in the limit as $\text{SNR} \rightarrow 0$ (due to the concavity of the throughput), wideband slope has emerged as a tool that enables us to analyze the energy efficiency at low but nonzero SNR levels. In our setup, we define energy per bit as

$$\frac{E_b}{N_0} = \frac{\text{SNR}}{r_{\text{avg}}^*(\text{SNR}, \theta)/m} \quad (53)$$

where the normalization with m is due to our assumption that r_{avg}^* is in the units of bits per m symbols (or equivalently per block).

The minimum energy per bit $\frac{E_b}{N_{0 \min}}$ under QoS constraints can be obtained from

$$\frac{E_b}{N_{0 \min}} = \lim_{\text{SNR} \rightarrow 0} \frac{\text{SNR}}{r_{\text{avg}}^*(\text{SNR}, \theta)/m} = \frac{1}{\dot{r}_{\text{avg}}^*(0)/m}. \quad (54)$$

At $\frac{E_b}{N_{0 \min}}$, the slope S_0 of the throughput versus E_b/N_0 (in dB) curve is defined as [41]

$$S_0 = \lim_{\frac{E_b}{N_0} \downarrow \frac{E_b}{N_{0 \min}}} \frac{r_{\text{avg}}^*(\text{SNR}, \theta)/m}{10 \log_{10} \frac{E_b}{N_0} - 10 \log_{10} \frac{E_b}{N_{0 \min}}} 10 \log_{10} 2. \quad (55)$$

The wideband slope can also be found from

$$S_0 = -\frac{2(\dot{r}_{\text{avg}}^*(0)/m)^2}{\ddot{r}_{\text{avg}}^*(0)/m} \log_e 2 \quad (56)$$

where $\dot{r}_{\text{avg}}^*(0)$ and $\ddot{r}_{\text{avg}}^*(0)$ are the first and second derivatives, respectively, of the function $r_{\text{avg}}^*(0)$ with respect to SNR at zero SNR. $\frac{E_b}{N_{0 \min}}$ and S_0 essentially provide a linear approximation of the throughput curve at low SNR levels.

B. Energy Efficiency with Constant Arrival Rate

In this section, we assume that the source arrival rate is fixed. Hence, we investigate the energy efficiency in the absence of source randomness and examine the impact of fading correlation and queueing constraints. As discussed in the previous section,

effective capacity, $C_E(\text{SNR}, \theta)$, characterizes the maximum constant arrival rate in the presence of QoS constraints described by the QoS exponent θ . Hence, we in this case have $r_{\text{avg}}^*(\text{SNR}, \theta) = C_E(\text{SNR}, \theta)$. In the following result, we provide the minimum bit energy and wideband slope expressions under these assumptions.

Theorem 10: Assume that the source arrival rate is constant. Then, the minimum energy per bit and wideband slope expressions as a function of the QoS exponent θ are given, respectively, by

$$\frac{E_b}{N_{0 \min}} = \frac{\log_e 2}{\mathbb{E}\{z\}} \quad (57)$$

and

$$S_0 = \frac{2(\mathbb{E}\{z\})^2}{\frac{\theta}{m \log_e 2} \sum_{i,j=1}^m \text{cov}\{z_i, z_j\} + \mathbb{E}\{z^2\}} \quad (58)$$

where $\text{cov}(z_i, z_j) = \mathbb{E}\{z_i z_j\} - \mathbb{E}\{z_i\}\mathbb{E}\{z_j\}$ is the covariance of z_i and z_j .

Proof: See Appendix H.

Remark 2: As can be seen in (57), the minimum energy per bit, which is achieved in the asymptotic regime in which SNR vanishes, does not depend on the QoS exponent θ , hence is not affected by the presence of the buffer limitations. Indeed, this is the fundamental limit in Gaussian channels [41]. Wideband slope S_0 , on the other hand, depends on the QoS constraints via the QoS exponent θ . It can be easily seen that higher the value of θ , the stricter the QoS constraints are and the smaller the value of the wideband slope is, indicating the increased energy requirements. Furthermore, it can be readily verified that wideband slope decreases with increased fading correlation. Or conversely, variations in the channel conditions are favorable for improved energy efficiency.

In Fig. 14, we plot the normalized maximum average arrival rate $\frac{1}{m} r_{\text{avg}}^*$ as a function of the energy per bit $\frac{E_b}{N_0}$ for different correlation factors ρ when $\theta = 1$ and $\mathbb{E}\{z\} = \mathbb{E}\{|h|^2\} = \sigma_h^2 = 1$. As predicted by Theorem 10, all curves converge to the same minimum energy per bit of $\frac{E_b}{N_{0 \min}} = \frac{\log_e 2}{\mathbb{E}\{z\}} = \log_e 2 = -1.59$ dB as SNR and hence r_{avg}^* vanish. On the other hand, wideband slopes are different for different values of ρ . As discussed above, as ρ and hence correlation diminishes from 1 to 0, slopes increase progressively. It is also interesting to note that in the absence of QoS constraints, i.e., when $\theta = 0$, such a distinction disappears. The wideband slope becomes $S_0 = \frac{2(\mathbb{E}\{z\})^2}{\mathbb{E}\{z^2\}}$, which clearly does not depend on the fading correlation.

C. Energy Efficiency with Discrete-Time Markov Sources

Starting with this subsection, we incorporate random arrivals into our energy efficiency analysis and determine how source randomness affects the performance.

1) *ON-OFF Discrete-Time Markov Sources:* We assume that data arrival is either in the ON or OFF state in each block duration of m symbols. As we have previously stated, in the ON state, λ bits arrive (i.e., the arrival rate is λ bits/block) while there are no arrivals in the OFF state. Below, we provide our results on energy efficiency.

Theorem 11: Assume that the source arrival rate is random and follows the described discrete-time ON-OFF Markov model.

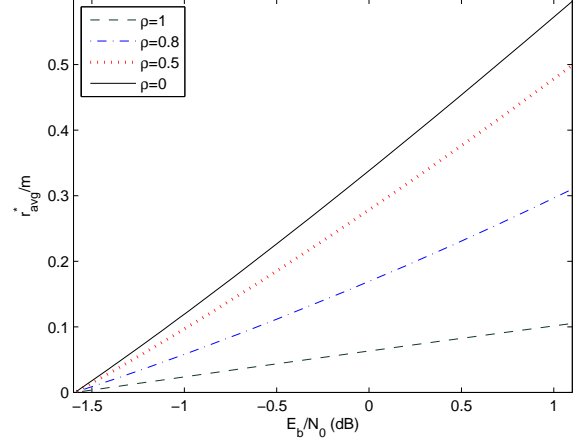


Fig. 14. Normalized maximum average arrival rate $\frac{1}{m} r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ for different fading correlations. $\theta = 1$, $m = 10$.

Then, the minimum energy per bit and wideband slope expressions as a function of the QoS exponent θ are given, respectively, by

$$\frac{E_b}{N_{0 \min}} = \frac{\log_e 2}{\mathbb{E}\{z\}} \quad (59)$$

and

$$S_0 = \frac{2(\mathbb{E}\{z\})^2}{\eta \frac{\theta m}{\log_e 2} (\mathbb{E}\{z\})^2 + \frac{\theta}{m \log_e 2} \sum_{i,j=1}^m \text{cov}\{z_i, z_j\} + \mathbb{E}\{z^2\}} \quad (60)$$

where η is defined in (37).

Proof: See Appendix I.

Interestingly, $\frac{E_b}{N_{0 \min}}$ again turns out to be a very robust quantity. Regardless of the buffer constraints, channel correlations, and randomness of the arrivals, the minimum received energy per bit is $\frac{E_b}{N_{0 \min}} = \log_e 2 = -1.59$ dB when $\mathbb{E}\{z\} = 1$. On the other hand, the impact of random arrivals on the wideband slope is perspicuous in (60). When compared to (58), we immediately notice that having random arrivals leads to the introduction of the term $\eta \frac{\theta m}{\log_e 2} (\mathbb{E}\{z\})^2$ in the denominator of (60). Notice that when $p_{22} = 1$ and $p_{11} = 0$, we have $P_{\text{ON}} = 1$, meaning that we have a constant arrival rate. In this case, $\eta = 0$ and indeed (60) specializes to (58). More generally, we have $\eta \geq 0$ for all $p_{11}, p_{22} \in [0, 1]$. Therefore, random arrivals potentially decreases the wideband slope and increases the energy requirements.

This is more clearly seen again in the special case in which $p_{11} = 1 - s$ and $p_{22} = s$. Now, we have $\eta = \frac{p_{11}}{p_{22}} = \frac{1-s}{s}$ and the wideband slope is

$$S_0 = \frac{2(\mathbb{E}\{z\})^2}{\frac{1-s}{s} \frac{\theta m}{\log_e 2} (\mathbb{E}\{z\})^2 + \frac{\theta}{m \log_e 2} \sum_{i,j=1}^m \text{cov}\{z_i, z_j\} + \mathbb{E}\{z^2\}} \quad (61)$$

As $P_{\text{ON}} = s$ decreases, the wideband slope decreases as well. Therefore, the source becoming more bursty leads to increased energy per bit. This is illustrated in Fig. 15 where maximum average arrival rate vs. energy per bit is plotted and the same

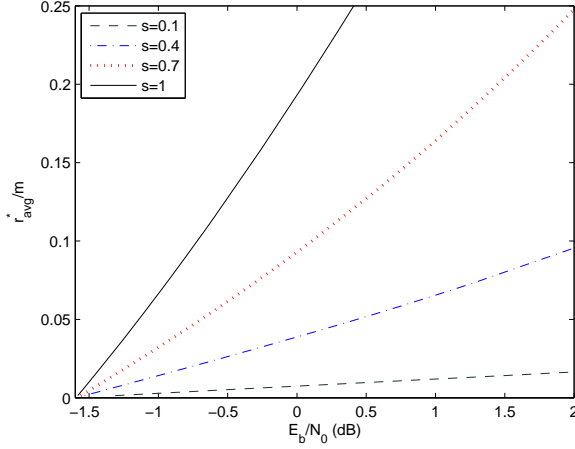


Fig. 15. Maximum average arrival rate $\frac{1}{m}r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ for different values of $P_{\text{ON}} = s$ when $\rho = 0.75$, $\theta = 1$, and $m = 10$.

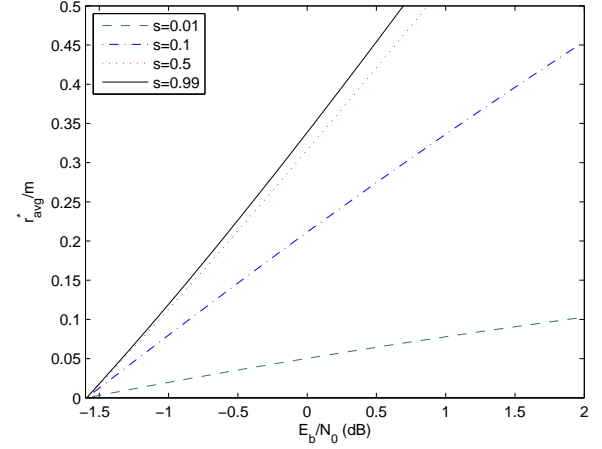


Fig. 16. Maximum average arrival rate $\frac{1}{m}r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ for different values of s when channel blocks are uncorrelated and $\theta = 1$. Number of states for the arrival process is $n = 10$.

channel fading and correlation model as in Fig. 14 is used. In this figure, we assume $\theta = 1$ and $\rho = 0.75$. As predicted, the minimum bit energies are all the same. However, we have diminishing slopes with decreasing P_{ON} . Note that for a fixed average arrival rate, as P_{ON} gets smaller, source becomes more bursty. Data arrives less frequently but with a higher rate. This in turn increases energy per bit as implied by smaller wideband slopes.

2) *Discrete-Time Markov Sources with n States*: In this model, we assume that there are $n - 1$ sources, each having its own ON and OFF states. In the ON state, a source sends data to the buffer at the rate of λ bits/block. Otherwise, it is in OFF state in which no data is generated. In this set-up, depending on how many sources are active (i.e., are in ON state), data arrivals to the buffer can be regarded as a discrete-time Markov process with n states. In the i^{th} state of this model, $(i - 1)$ sources are active. For simplicity, we assume that the probability of each source being active in a given block is s , independent of the previous states and of the other sources. Then, the state probabilities will be given by

$$\pi_i = \binom{n-1}{i-1} s^{i-1} (1-s)^{n-i} \text{ for } i = 1, 2, \dots, n. \quad (62)$$

Note that the system is essentially memoryless because each state is independent of the previous state. Hence, transition probability matrix becomes

$$\mathbf{J} = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \\ \pi_1 & \pi_2 & \cdots & \pi_n \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{bmatrix}. \quad (63)$$

Using (62) we can write the average rate expression as

$$r_{\text{avg}} = \sum_{i=1}^n \binom{n-1}{i-1} q^{i-1} (1-q)^{n-i} (i-1)\lambda = (n-1)q\lambda. \quad (64)$$

For this case, we do not have closed-form expressions. However, we can easily obtain the effective bandwidth and maximum

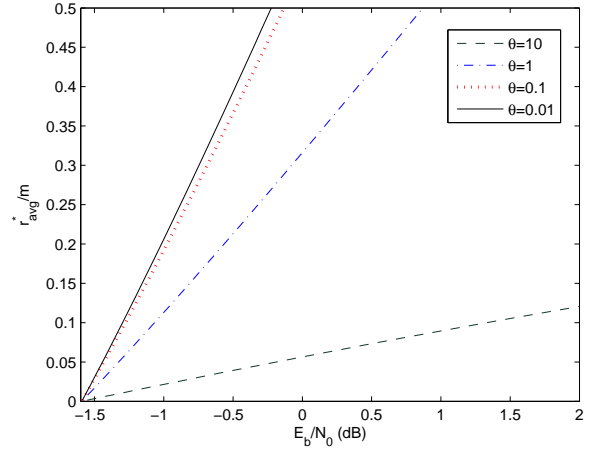


Fig. 17. Maximum average arrival rate $\frac{1}{m}r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ for different θ values when channel blocks are uncorrelated and $s = 0.5$. Number of states for arrival process is $n = 10$.

average arrival rate numerically. In particular, by numerically solving (33) and using (64), we can determine the maximum average arrival rate r_{avg}^* as a function of SNR. In Fig. 16, we display the maximum average arrival rate as a function of energy per bit. Similarly as in the simple ON/OFF model, we observe that when source burstiness is decreased by increasing s , energy efficiency improves.

To have a better understanding of the effect of the QoS constraints, average arrival rate curves in Fig. 17 are obtained for different θ values. We first notice that QoS exponent θ does not have any effect on the minimum energy per bit because all curves merge at -1.59 dB which is again the minimum energy per bit for all θ values. However, energy efficiency degrades with stricter QoS conditions as increasing θ reduces the wideband slope.

Finally, for comparison purposes, we depict the throughput as a function of SNR in Fig. 18 for different source characteristics and θ values. The trends in the throughput vs. SNR curves for

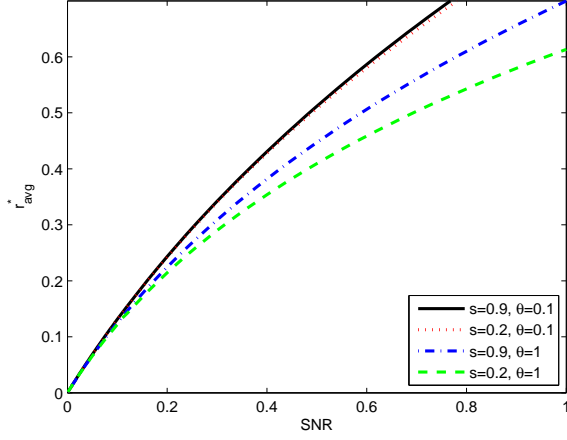


Fig. 18. Maximum average arrival rate r_{avg}^* vs. signal-to-noise ratio SNR when channel blocks are uncorrelated. Number of states of the arrival process is $n = 10$.

the considered n -state discrete Markov source are observed to be similar to those in Fig. 3 plotted for the ON-OFF discrete Markov source. For instance, again increased source burstiness (i.e., lower values of s) and stricter queueing constraints (i.e., higher θ values) result in the degradation of the throughput for both two-state (ON/OFF) and n -state source models.

D. Energy Efficiency with Markov Fluid Sources

1) *ON-OFF Markov Fluid Model*: Now, we consider Markov fluid sources with two states, namely OFF state with no arrivals and ON state in which the arrival rate is λ . The generating matrix is defined in (24). Minimum energy per bit and wideband slope are derived in the following result.

Theorem 12: Assume that the source arrival is modeled by a two-state (ON-OFF) continuous-time Markov chain. Then, the minimum energy per bit and wideband slope expressions as a function of the QoS exponent θ are given, respectively, by

$$\frac{E_b}{N_0 \min} = \frac{\log_e 2}{\mathbb{E}\{z\}} \quad (65)$$

and

$$\mathcal{S}_0 = \frac{2(\mathbb{E}\{z\})^2}{\zeta \frac{\theta m}{\log_e 2} (\mathbb{E}\{z\})^2 + \frac{\theta}{m \log_e 2} \sum_{i,j=1}^m \text{cov}\{z_i, z_j\} + \mathbb{E}\{z^2\}} \quad (66)$$

where $\zeta = \frac{2\beta}{\alpha(\alpha+\beta)}$ as defined in (45).

Proof: See Appendix J.

Similarly as before, QoS constraints and source randomness do not affect the minimum energy per bit. On the other hand, it is seen in (66) that the impact of source arrival characteristics on the wideband slope is via the state transition rates α and β . For instance, larger the α value, the higher the wideband slope is. This is due to the fact that as α , which is the transition rate from OFF state to ON state, increases, the system is more likely to be in the ON state. Contrarily, wideband slope diminishes with increasing

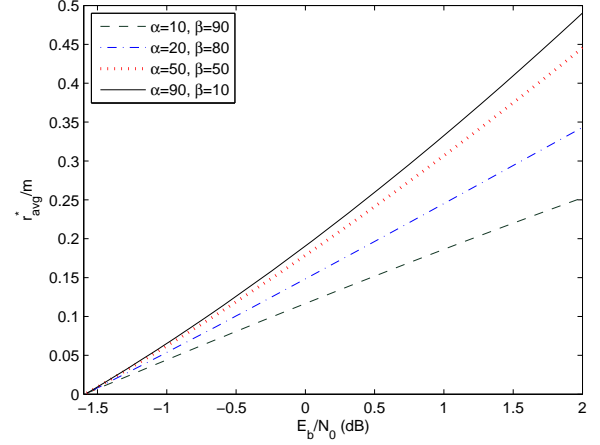


Fig. 19. Maximum average arrival rate $\frac{1}{m} r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ when $\rho = 0.75$, $\theta = 1$, and $m = 10$.

β . This is expected as well since larger β leads to higher OFF-state probabilities. The effect of α and β is illustrated in Fig. 19, where maximum average arrival rate vs. energy per bit is plotted. In this figure, we set $\theta = 1$ and $\rho = 0.75$. As predicted, the same minimum energy per bit is achieved for different values of α and β , while wideband slope increases with increasing α or decreasing β .

2) *n-State Markov Fluid Birth-Death Process*: In this subsection, we consider a birth-death process for the Markov fluid source. We assume that there are n states and the arrival rate in the i^{th} state is $(i-1)\lambda$ for $i = 1, \dots, n$. The generating matrix for the birth-death process is in the form of

$$\mathbf{G} = \begin{bmatrix} -\alpha & \alpha & 0 & \dots & \dots & 0 \\ \beta & -(\alpha + \beta) & \alpha & 0 & \dots & 0 \\ 0 & & \ddots & & & \vdots \\ \vdots & & & \ddots & & 0 \\ 0 & \dots & 0 & \beta & -(\alpha + \beta) & \alpha \\ 0 & \dots & \dots & 0 & -\beta & \beta \end{bmatrix}. \quad (67)$$

Hence, the transition rate from state i to state $i+1$ is α whereas the transition rate from state i to state $i-1$ is β . The effective bandwidth of this source, which does not have a simple closed-form expression, can be found from (22). In order to conduct an energy efficiency analysis, average arrival rate needs to be identified as well. Using (23) and (67), we can easily determine that the stationary distribution as

$$\pi_i = \frac{\left(\frac{\alpha}{\beta}\right)^{i-1} - \left(\frac{\alpha}{\beta}\right)^i}{1 - \left(\frac{\alpha}{\beta}\right)^n} \text{ for } i = 1, 2, \dots, n \quad (68)$$

when $\frac{\alpha}{\beta} \neq 1$, $\alpha \neq 0$ and $\beta \neq 0$. If any of these inequalities is not satisfied, state probabilities can be obtained by limiting functions.

Now, under the assumptions that $\alpha \neq \beta$ and the arrival rate in

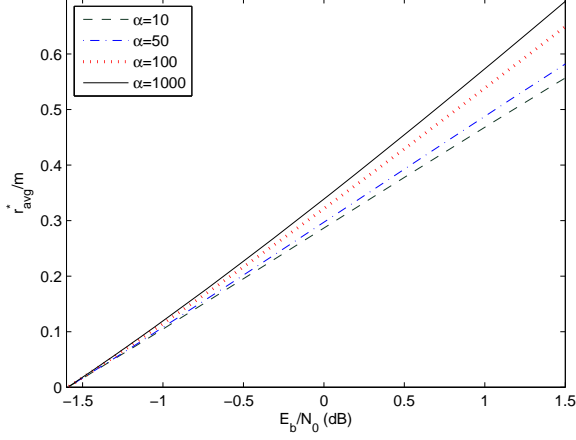


Fig. 20. Maximum average arrival rate $\frac{1}{m}r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ when channel blocks are uncorrelated. $\theta = 1$, $\beta = 100$ and the number of states of the arrival process is $n = 10$.

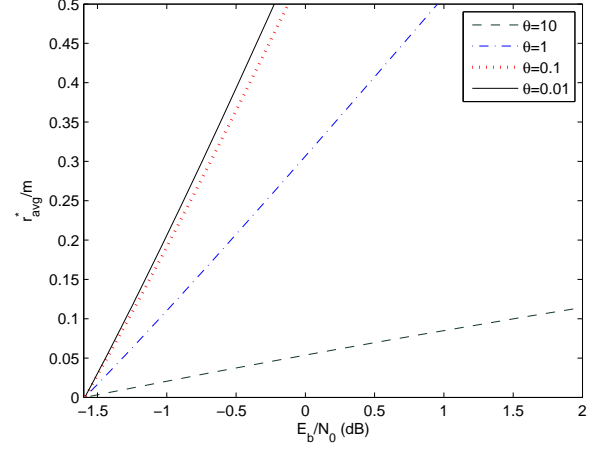


Fig. 21. Maximum average arrival rate r_{avg}^* vs. energy per bit $\frac{E_b}{N_0}$ when channel blocks are uncorrelated and $\alpha = \beta = 50$. Number of states of the arrival model is 10.

state i is $\lambda_i = (i-1)\lambda(\text{SNR})$, the average arrival rate is given by

$$r_{\text{avg}} = \frac{\xi(1 - n\xi^{n-1} + (n-1)\xi^n)}{(1-\xi)(1-\xi^n)}\lambda(\text{SNR}) \quad (69)$$

where $\xi = \frac{\alpha}{\beta}$.

Remark 3: Note that (69) specializes to (27) if $n = 2$. When $\xi \rightarrow \infty$, the probability of the n^{th} state approaches 1, and the arrival rate will be $(n-1)\lambda(\text{SNR})$ at steady state. On the other hand, for $\xi = 0$, the state of the source is stuck at the first state in which the arrival rate is zero.

Numerically, we can obtain the effective bandwidth of the n -state birth-death Markov fluid process using (22). Subsequently, solving (33) and incorporating (69), we can determine the maximum average arrival rate $r_{\text{avg}}^*(\text{SNR})$, which we further employ for characterizing the energy efficiency. The results of this numerical analysis are displayed in the following figures. In Fig. 20, we demonstrate the effect of α on the energy efficiency. In particular, when β is kept fixed, increasing α improves the energy efficiency as in the two-state case. We illustrate the effect of QoS constraints in Fig. 21. Similar conclusions as before readily apply. QoS exponent θ does not alter the minimum bit energy, which is -1.59 dB again, but the wideband slope is reduced with increasing θ .

E. Energy Efficiency with Markov-Modulated Poisson Process

1) *ON-OFF Markov-Modulated Poisson Process:* Again, we initially address the two-state model in which there are no arrivals in the OFF state and the intensity of the Poisson arrival process is λ in the ON state. The generating matrix \mathbf{G} is the same as in (24).

Theorem 13: Assume that the source arrival is modeled by a two-state (ON/OFF) Markov modulated Poisson process. Then, the minimum energy per bit and wideband slope expressions as a function of the QoS exponent θ are given, respectively, by

$$\frac{E_b}{N_{0\min}} = \frac{e^\theta - 1}{\theta} \frac{\log_e 2}{\mathbb{E}\{z\}} \quad (70)$$

and

$$\mathcal{S}_0 = \frac{\frac{2\theta}{e^\theta - 1}(\mathbb{E}\{z\})^2}{\zeta \frac{\theta m}{\log_e 2}(\mathbb{E}\{z\})^2 + \frac{\theta}{m \log_e 2} \sum_{i,j=1}^m \text{cov}\{z_i, z_j\} + \mathbb{E}\{z^2\}} \quad (71)$$

where $\zeta = \frac{2\beta}{\alpha(\alpha+\beta)}$ as defined in (45).

Proof: See Appendix K.

Remark 4: It is interesting to observe that, unlike the previous arrival models, minimum energy per bit in the case of MMPP source depends on the QoS exponent θ . More specifically, minimum energy per bit increases with $(e^\theta - 1)/\theta$ which is an increasing monotonic function of θ and always greater than one for $\theta > 0$. On the other hand, as $\theta \rightarrow 0$, $(e^\theta - 1)/\theta \rightarrow 1$. Therefore, $\frac{E_b}{N_{0\min}} \geq \frac{\log_e 2}{\mathbb{E}\{z\}}$ with equality only if no QoS constraints are imposed (i.e., when $\theta = 0$). Furthermore, in addition to its significant impact on the minimum energy per bit, increasing θ leads to much quicker reduction in the wideband slope due to the presence of the term $\frac{\theta}{e^\theta - 1}$ in (71). Hence, overall, energy costs grow very fast as θ increases. This is again because of the additional randomness arising from Poisson arrivals in the ON state.

Remark 5: From (71), we note that the effect of the state transition rates α and β on the energy efficiency is the same as in the Markov fluid source model. Increasing α or decreasing β improves the energy efficiency of the system because the burstiness of the data arrivals is reduced and the buffer overflows can be avoided at lower energy costs.

We plot the maximum average arrival rate vs. energy per bit in Fig. 22. We set $\mathbb{E}\{z\} = 1$ and $\theta = 1$ for which the minimum energy per bit is 0.76 dB. The increase in bit energy with respect to -1.59 dB is due to $10 \log_{10}((e^\theta - 1)/\theta) = 10 \log_{10}(e - 1)$ for $\theta = 1$. From the figure, we can again infer that adjusting α or β to increase the ON state probability makes the system more energy efficient due to the increase in the wideband slope.

2) *n-State Markov-Modulated Poisson Process:* Finally, we consider an n -state MMPP process and assume that the intensity

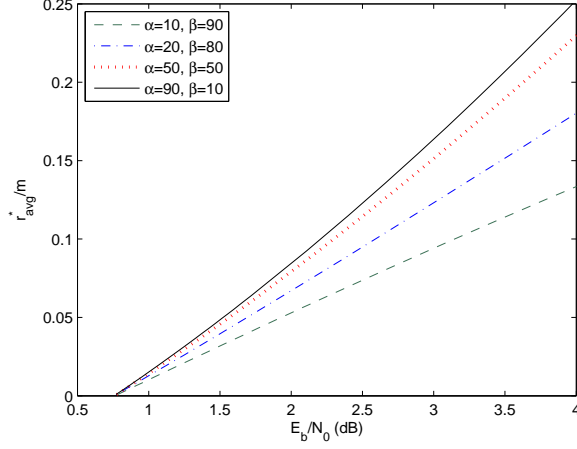


Fig. 22. Maximum average arrival rate $\frac{1}{m}r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ when $\rho = 0.75$, $\theta = 1$, and $m = 10$.

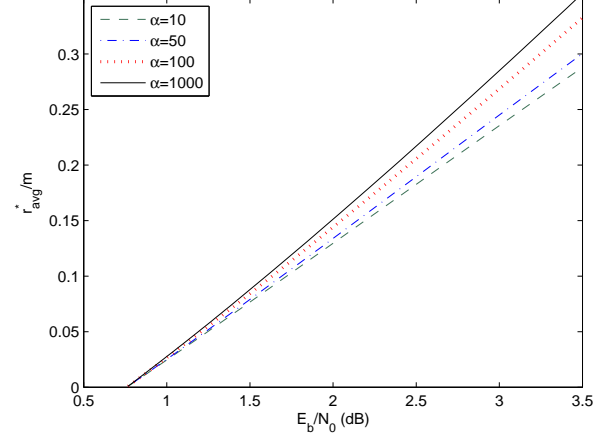


Fig. 23. Maximum average arrival rate $\frac{1}{m}r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ when channel blocks are uncorrelated. $\theta = 1$, $\beta = 100$ and the number of Markov states for the arrival process is $n = 10$.

of the Poisson arrivals in the i^{th} state is $(i-1)\lambda$. For the Markov transitions between states, we consider the birth-death process and adopt the transition rate matrix \mathbf{G} from (67). We solve for the maximum intensity $\lambda^*(\text{SNR}, \theta)$ by incorporating (28) into (33). Then, using the expression in (69), we obtain $r_{\text{avg}}^*(\text{SNR}, \theta)$.

In Figs. 23 and 24, we depict the maximum average arrival rate as a function of the energy per bit with uncorrelated channel coefficients being assumed in each block. In Fig. 23, we set $n = 10$, $\theta = 1$ and $\beta = 100$, and demonstrate how α influences the energy efficiency of system. The observation has similarities with other Markovian sources regarding the source burstiness. Interestingly, the minimum energy per bit is again 0.76 dB as in the two-state case, leading to the conclusion that the number of states does not alter $\frac{E_b}{N_0}_{\min}$ in this case. The degradation in energy efficiency due to increased θ is shown in Fig. 24. As described in the two-state case, higher values of θ (i.e., stricter QoS constraints) result in higher $\frac{E_b}{N_0}_{\min}$ and smaller wideband slope. Therefore, even for relatively small increases in θ , we can have large gaps between curves, indicating significantly high energy costs.

V. CONCLUSION

In this paper, we have studied the throughput and energy efficiency in wireless fading channels when data arrivals to the buffer are random and constraints on buffer overflow probabilities are imposed. We have considered discrete-time Markov, Markov fluid, and Markov-modulated Poisson (MMPP) sources, and formulated the maximum average arrival rates of these sources in closed-form in the special case in which the source has only ON and OFF states. We have shown that as the QoS exponent θ vanishes (i.e., no QoS constraints are imposed), throughput converges to the ergodic capacity and the performance becomes independent of the source characteristics and channel fading correlations. On the other hand, we have demonstrated for $\theta > 0$ that the statistics of the source arrivals and channel conditions have significant impact on the throughput. More specifically, it is seen that while fast channel variations (e.g., having fading less

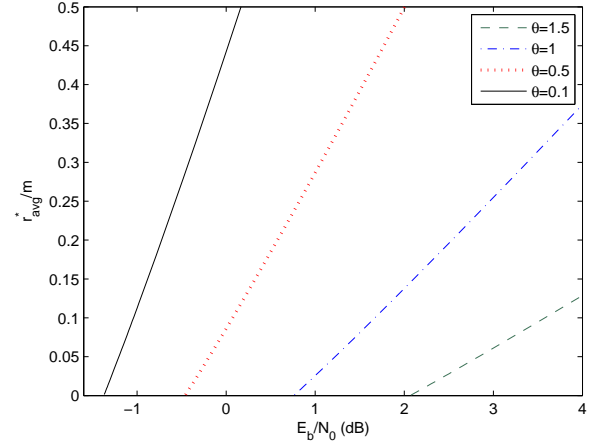


Fig. 24. Maximum average arrival rate $\frac{1}{m}r_{\text{avg}}^*$ vs. energy per bit $\frac{E_b}{N_0}$ when channel blocks are uncorrelated and $\alpha = \beta = 50$. Number of Markov states for the arrival process is $n = 10$.

correlated) can be beneficial, source randomness and burstiness degrade the throughput. In particular, performance loss can be severe with MMPP sources especially if θ is large. Similar conclusions are drawn in the energy efficiency analysis as well. We have shown that the minimum energy per bit is only a function of the marginal distribution of the fading for constant, discrete-time Markov and Markov fluid arrivals. However, for MMPP arrivals, this asymptotic measure of energy efficiency increases as the QoS exponent θ increases and hence as stricter queueing constraints are imposed. For practical scenarios, wideband slope is a better indicator of the performance and is always seen to depend on the source statistics, channel correlations, and QoS exponent θ . In particular, we have noted that wideband slope diminishes with increasing source burstiness and stricter QoS constraints, indicating increased energy costs in such cases. While this paper primarily concentrates on two-state Markovian sources, numerical results are provided in the energy efficiency analysis

for several simple n -state models with which we have similar observations as above.

APPENDIX

A. Proof of Theorem 1:

Using the effective bandwidth formulation in (19), we can express (33) in the following equivalent form:

$$\frac{1}{\theta} \log_e \left(\frac{p_{11} + p_{22} e^{\lambda\theta} + \sqrt{(p_{11} + p_{22} e^{\lambda\theta})^2 - 4(p_{11} + p_{22} - 1)e^{\lambda\theta}}}{2} \right) = C_E. \quad (72)$$

Then, we rewrite the above equality as

$$p_{11} + p_{22} e^{\lambda\theta} + \sqrt{(p_{11} + p_{22} e^{\lambda\theta})^2 - 4(p_{11} + p_{22} - 1)e^{\lambda\theta}} = 2e^{\theta C_E}, \quad (73)$$

from which, after moving the first two terms on the left-hand side to the right-hand side and taking the square of both sides, we obtain

$$(p_{11} + p_{22} e^{\lambda\theta})^2 - 4(p_{11} + p_{22} - 1)e^{\lambda\theta} = (2e^{\theta C_E} - p_{11} - p_{22} e^{\lambda\theta})^2. \quad (74)$$

Now, by simply exchanging the second term on the left-hand side with the term on the right-hand side, we have

$$(p_{11} + p_{22} e^{\lambda\theta})^2 - (2e^{\theta C_E} - p_{11} - p_{22} e^{\lambda\theta})^2 = 4(p_{11} + p_{22} - 1)e^{\lambda\theta}, \quad (75)$$

$$(2p_{11} + 2p_{22} e^{\lambda\theta} - 2e^{\theta C_E}) 2e^{\theta C_E} = 4(p_{11} + p_{22} - 1)e^{\lambda\theta}. \quad (76)$$

After further rearrangements, we have

$$(p_{11} + p_{22} - 1 - p_{22} e^{\theta C_E}) e^{\lambda\theta} = p_{11} e^{\theta C_E} - e^{2\theta C_E}. \quad (77)$$

Solving the equation for λ , we get

$$\lambda^*(\text{SNR}, \theta) = \frac{1}{\theta} \log_e \left(\frac{e^{2\theta C_E(\text{SNR}, \theta)} - p_{11} e^{\theta C_E(\text{SNR}, \theta)}}{1 - p_{11} - p_{22} + p_{22} e^{\theta C_E(\text{SNR}, \theta)}} \right) \quad (78)$$

which provides the maximum arrival rate in the ON state. We can now express the maximum arrival rate as $r^*(\text{SNR}, \theta) = P_{\text{ON}} \lambda^*(\text{SNR}, \theta)$ and obtain the expression in (34). \square

B. Proof of Theorem 2:

Let us define

$$\psi(\theta) = e^{-\theta C_E(\text{SNR}, \theta)} = \mathbb{E} \left\{ e^{-\frac{\theta}{\log_e 2} \sum_{i=1}^m \log_e(1 + \text{SNR} z_i)} \right\}. \quad (79)$$

The following properties of ψ can be verified easily:

$$\psi(0) = 1, \quad (80)$$

$$\dot{\psi}(\theta) = \mathbb{E} \left\{ - \sum_{i=1}^m \log_2(1 + \text{SNR} z_i) e^{-\theta \sum_{i=1}^m \log_2(1 + \text{SNR} z_i)} \right\}, \quad (81)$$

$$\ddot{\psi}(\theta) = \mathbb{E} \left\{ \left(\sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \right)^2 e^{-\theta \sum_{i=1}^m \log_2(1 + \text{SNR} z_i)} \right\}, \quad (82)$$

and

$$\dot{\psi}(0) = - \mathbb{E} \left\{ \sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \right\}, \quad (83)$$

$$\ddot{\psi}(0) = \mathbb{E} \left\{ \left(\sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \right)^2 \right\} \quad (84)$$

where $\dot{\psi}$ and $\ddot{\psi}$ denote the first and second derivatives of ψ with respect to θ , respectively. Additionally, we define $r_{\text{avg}}^*(\text{SNR}, \theta)$ as

$$r_{\text{avg}}^*(\text{SNR}, \theta) = \frac{f_1(\theta)}{\theta}. \quad (85)$$

Therefore, by applying L'Hopital's rule and letting $\theta \rightarrow 0$, maximum average arrival rate and its slope can be easily found as

$$\lim_{\theta \rightarrow 0} r_{\text{avg}}^*(\text{SNR}, \theta) = \dot{f}_1(0), \quad (86)$$

$$\frac{\partial r_{\text{avg}}^*(\text{SNR}, \theta)}{\partial \theta} \Big|_{\theta=0} = \frac{\ddot{f}_1(0)}{2}. \quad (87)$$

Now, replacing $e^{-\theta C_E(\text{SNR}, \theta)}$ with $\psi(\theta)$ in the expression of $r_{\text{avg}}^*(\text{SNR}, \theta)$ in (34), we can express $f_1(\theta)$ as

$$f_1(\theta) = P_{\text{ON}} [\log_e(1 - p_{11}\psi(\theta)) - \log_e((1 - p_{11} - p_{22})\psi^2(\theta) + p_{22}\psi(\theta))]. \quad (88)$$

Therefore, we derive $\dot{f}_1(\theta)$ when $\theta \rightarrow 0$ as

$$\dot{f}_1(0) = P_{\text{ON}} \left[\frac{-p_{11}\dot{\psi}(0)}{1 - p_{11}\psi(0)} - \frac{(1 - p_{11} - p_{22})\dot{\psi}(0)}{(1 - p_{11} - p_{22})\psi(0) + p_{22}} - \frac{\dot{\psi}(0)}{\psi(0)} \right] \quad (89)$$

$$= P_{\text{ON}} \left[\frac{-p_{11}}{1 - p_{11}} - \frac{1 - p_{11} - p_{22}}{1 - p_{11}} - 1 \right] \dot{\psi}(0) \quad (90)$$

$$= \mathbb{E} \left\{ \sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \right\}. \quad (91)$$

Note that (89) follows by taking the first derivative of the expression in (88) with respect to θ , and (90) is obtained using the property that $\psi(0) = 1$. Finally, (91) and hence the result in (35) immediately follow from (20), (83) and (86).

Next, we determine the slope of the throughput in (36) as the QoS exponent θ approaches zero. For this, we only need to derive the second derivative expression $\ddot{f}_1(0)$, which is done at the top of the next page. (92), (93) and (94) follow from straightforward algebraic steps. Inserting (83) and (84) into (94), we obtain (95). Finally, the result in (36) follows by combining (95) and (87). \square

C. Proof of Theorem 3:

In the analysis of the high-SNR slope of the effective capacity, it has been shown in [28] that

$$\begin{aligned} & -\frac{1}{\theta} \log_e \mathbb{E} \left\{ e^{-\theta \log_2(1 + \text{SNR} z)} \right\} \\ &= \begin{cases} \frac{1}{\theta \log_2 e} \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } \theta > \frac{1}{\log_2 e} \\ \log_2 \text{SNR} + \mathcal{O}(1) & \text{if } 0 < \theta < \frac{1}{\log_2 e} \end{cases} \quad (96) \end{aligned}$$

$$\ddot{f}_1(0) = P_{\text{ON}} \frac{d}{d\theta} \left[\frac{-p_{11}\dot{\psi}(0)}{1-p_{11}\psi(0)} - \frac{(1-p_{11}-p_{22})\dot{\psi}(0)}{(1-p_{11}-p_{22})\psi(0)+p_{22}} - \frac{\dot{\psi}(0)}{\psi(0)} \right] \Big|_{\theta=0} \quad (92)$$

$$= P_{\text{ON}} \left\{ \left[\frac{-p_{11}}{1-p_{11}} - \frac{1-p_{11}-p_{22}}{1-p_{11}} - 1 \right] \ddot{\psi}(0) + \left[-\frac{p_{11}^2}{(1-p_{11})^2} + \frac{(1-p_{11}-p_{22})^2}{(1-p_{11})^2} + 1 \right] (\dot{\psi}(0))^2 \right\} \quad (93)$$

$$= -\ddot{\psi}(0) + (1-\eta) (\dot{\psi}(0))^2 \quad (94)$$

$$= -\mathbb{E} \left\{ \left(\sum_{i=1}^m \log_2(1 + \text{SNR}z_i) \right)^2 \right\} + (1-\eta) \left[\mathbb{E} \left\{ \sum_{i=1}^m \log_2(1 + \text{SNR}z_i) \right\} \right]^2 \quad (95)$$

where z is exponentially distributed with unit mean. If we assume that fading in each block is i.i.d., then the effective capacity expression in (13) becomes

$$C_E(\text{SNR}, \theta) = -\frac{1}{\theta} \log_e \mathbb{E} \left\{ e^{-\theta \sum_{i=1}^m \log(1 + \text{SNR}z_i)} \right\} \quad (97)$$

$$= -\frac{1}{\theta} \log_e \left(\prod_{i=1}^m \mathbb{E} \left\{ e^{-\theta \log(1 + \text{SNR}z_i)} \right\} \right) \quad (98)$$

$$= -\frac{m}{\theta} \log_e \mathbb{E} \left\{ e^{-\theta \log(1 + \text{SNR}z)} \right\}. \quad (99)$$

Furthermore, the maximum average arrival rate in (34) can be expressed as

$$r_{\text{avg}}^*(\text{SNR}, \theta) = \frac{P_{\text{ON}}}{\theta} \log_e \left(\frac{e^{2\theta C_E(\text{SNR}, \theta)} (1 - p_{11}e^{-\theta C_E(\text{SNR}, \theta)})}{e^{\theta C_E(\text{SNR}, \theta)} ((1 - p_{11} - p_{22})e^{-\theta C_E(\text{SNR}, \theta)} + p_{22})} \right) \quad (100)$$

$$= \frac{P_{\text{ON}}}{\theta} \log_e \left(\frac{e^{\theta C_E(\text{SNR}, \theta)} (1 - p_{11}e^{-\theta C_E(\text{SNR}, \theta)})}{(1 - p_{11} - p_{22})e^{-\theta C_E(\text{SNR}, \theta)} + p_{22}} \right) \quad (101)$$

$$= \frac{P_{\text{ON}}}{\theta} \left(\log_e e^{\theta C_E(\text{SNR}, \theta)} + \log_e (1 - p_{11}e^{-\theta C_E(\text{SNR}, \theta)}) - \log_e ((1 - p_{11} - p_{22})e^{-\theta C_E(\text{SNR}, \theta)} + p_{22}) \right) \quad (102)$$

$$= \frac{P_{\text{ON}}}{\theta} \log_e e^{\theta C_E(\text{SNR}, \theta)} + \mathcal{O}(1) \quad (103)$$

$$= P_{\text{ON}} C_E(\text{SNR}, \theta) + \mathcal{O}(1) \quad (104)$$

where (101) and (102) follow from straightforward algebraic operations and (103) is due to the fact that $C_E(\text{SNR}, \theta)$ increases without bound as SNR increases and hence the term $e^{-\theta C_E(\text{SNR}, \theta)}$ vanishes asymptotically in the formulations.

Finally, combining (96), (99), and (104), we immediately obtain the desired result in (38) for the cases in which $\theta > 0$. When $\theta = 0$, the result follows from (35) in Theorem 2. \square

D. Proof of Theorem 4:

Using (25), we can rewrite (33) as

$$(\theta\lambda - (\alpha + \beta) - 2\theta C_E)^2 = (\theta\lambda - (\alpha + \beta))^2 + 4\alpha\theta\lambda \quad (105)$$

which can further be simplified to

$$-2\theta C_E(2\theta\lambda - 2(\alpha + \beta) - 2\theta C_E) = 4\alpha\theta\lambda. \quad (106)$$

Next, solving for λ , we obtain

$$\lambda^*(\text{SNR}, \theta) = \frac{\theta C_E(\text{SNR}, \theta) + \alpha + \beta}{\theta C_E(\text{SNR}, \theta) + \alpha} C_E(\text{SNR}, \theta). \quad (107)$$

Finally, using the expression in (27), we derive the maximum average arrival rate given in (42). \square

E. Proof of Theorem 5:

Similar as in the Proof of Theorem 2 in Appendix B, we define $r_{\text{avg}}^*(\text{SNR}, \theta) = \frac{f_2(\theta)}{\theta}$ with

$$f_2(\theta) = -P_{\text{ON}} \frac{\alpha + \beta - \log_e \psi(\theta)}{\alpha - \log_e \psi(\theta)} \log_e \psi(\theta). \quad (108)$$

Now, we have (86) and (87) hold with f_1 replaced with f_2 . The remainder of the proof requires only the determination of the first and second derivatives of $f_2(\theta)$ at $\theta = 0$. The first derivative $\dot{f}_2(0)$ is given at the top of the next page in (109)-(112). Note that (109) and (110) follow from straightforward algebraic steps, and (111) is obtained by noting the property that $\psi(0) = 1$. Finally, (112) and hence the result in (43) immediately follow from (26), (83) and (86).

Next, we obtain the slope expression in (44) in the limit as the QoS exponent θ approaches zero. For this, we characterize the second derivative expression $\ddot{f}_2(0)$ on the next page in (113)-(116). (113), (114) are readily obtained and (115) is determined by noting that $\psi(0) = 1$. We incorporate (83) and (84) into (115) to obtain (116). The result in (44) follows by combining (116) and (87) (with f_1 replaced with f_2). \square

F. Proof of Theorem 7:

We find the maximum average arrival rate $r_{\text{avg}}^*(\text{SNR}, \theta)$ by incorporating (29) into (33) and expressing (33) as

$$\begin{aligned} & ((e^\theta - 1)\lambda - (\alpha + \beta) - 2\theta C_E)^2 \\ &= ((e^\theta - 1)\lambda - (\alpha + \beta))^2 + 4\alpha(e^\theta - 1)\lambda. \end{aligned} \quad (117)$$

Similarly as in the proof of Theorem 4, we can simplify the above equality and solve for the maximum Poisson arrival intensity in the ON state to obtain

$$\lambda^*(\text{SNR}, \theta) = \frac{\theta [\theta C_E(\text{SNR}, \theta) + \alpha + \beta]}{(e^\theta - 1) [\theta C_E(\text{SNR}, \theta) + \alpha]} C_E(\text{SNR}, \theta). \quad (118)$$

With this characterization, the maximum average arrival rate is readily obtained from (30). \square

$$\dot{f}_2(0) = \lim_{\theta \rightarrow 0} -P_{\text{ON}} \left\{ \left[\frac{-\frac{\dot{\psi}(\theta)}{\psi(\theta)}}{\alpha - \log_e \psi(\theta)} - \frac{(\alpha + \beta - \log_e \psi(\theta)) \left(-\frac{\dot{\psi}(\theta)}{\psi(\theta)}\right)}{(\alpha - \log_e \psi(\theta))^2} \right] \log_e \psi(\theta) + \frac{\alpha + \beta - \log_e \psi(\theta)}{\alpha - \log_e \psi(\theta)} \frac{\dot{\psi}(\theta)}{\psi(\theta)} \right\} \quad (109)$$

$$= \lim_{\theta \rightarrow 0} -P_{\text{ON}} \left[1 + \frac{\alpha\beta}{(\alpha - \log_e \psi(\theta))^2} \right] \frac{\dot{\psi}(\theta)}{\psi(\theta)} \quad (110)$$

$$= -P_{\text{ON}} \frac{\alpha + \beta}{\alpha} \dot{\psi}(0) \quad (111)$$

$$= \mathbb{E} \left\{ \sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \right\} \quad (112)$$

$$\ddot{f}_2(0) = \lim_{\theta \rightarrow 0} -P_{\text{ON}} \frac{d}{d\theta} \left\{ \left[1 + \frac{\alpha\beta}{(\alpha - \log_e \psi(\theta))^2} \right] \frac{\dot{\psi}(\theta)}{\psi(\theta)} \right\} \quad (113)$$

$$= \lim_{\theta \rightarrow 0} -P_{\text{ON}} \frac{d}{d\theta} \left\{ \frac{-2\alpha\beta \left(-\frac{\dot{\psi}(\theta)}{\psi(\theta)}\right)}{(\alpha - \log_e \psi(\theta))^3} \left(\frac{\dot{\psi}(\theta)}{\psi(\theta)}\right) + \left[1 + \frac{\alpha\beta}{(\alpha - \log_e \psi(\theta))^2} \right] \left[\frac{\ddot{\psi}(\theta)}{\psi(\theta)} - \left(\frac{\dot{\psi}(\theta)}{\psi(\theta)}\right)^2 \right] \right\} \quad (114)$$

$$= -\ddot{\psi}(0) + \left(1 - \frac{2\beta}{\alpha(\alpha + \beta)}\right) (\dot{\psi}(0))^2 \quad (115)$$

$$= -\mathbb{E} \left\{ \left(\sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \right)^2 \right\} + \left(1 - \frac{2\beta}{\alpha(\alpha + \beta)}\right) \left[\mathbb{E} \left\{ \sum_{i=1}^m \log_2(1 + \text{SNR} z_i) \right\} \right]^2 \quad (116)$$

G. Proof of Theorem 8:

Employing $f_2(\theta)$ defined in (108), we can express the maximum average arrival rate as

$$r_{\text{avg}}^*(\text{SNR}, \theta) = \frac{f_2(\theta)}{e^\theta - 1}. \quad (119)$$

Then, the throughput in the limit as θ approaches zero is given by

$$\lim_{\theta \rightarrow 0} r_{\text{avg}}^*(\text{SNR}, \theta) = \lim_{\theta \rightarrow 0} \frac{\dot{f}_2(\theta)}{e^\theta} = \dot{f}_2(0). \quad (120)$$

Inserting the result from (112) into (120), we obtain (49). Next, we determine the slope of the throughput when θ approaches zero:

$$\left. \frac{\partial r_{\text{avg}}^*(\text{SNR}, \theta)}{\partial \theta} \right|_{\theta=0} = \lim_{\theta \rightarrow 0} \frac{\dot{f}_2(\theta)}{e^\theta - 1} - \frac{e^\theta f_2(\theta)}{(e^\theta - 1)^2} \quad (121)$$

$$= \lim_{\theta \rightarrow 0} \frac{(e^\theta - 1) \dot{f}_2(\theta) - e^\theta f_2(\theta)}{(e^\theta - 1)^2} \quad (122)$$

$$= \lim_{\theta \rightarrow 0} \frac{(e^\theta - 1) \ddot{f}_2(\theta) - e^\theta f_2(\theta)}{2(e^\theta - 1)e^\theta} \quad (123)$$

$$= \frac{\ddot{f}_2(0)}{2} - \frac{1}{2} \lim_{\theta \rightarrow 0} \frac{f_2(\theta)}{e^\theta - 1} \quad (124)$$

$$= \frac{\ddot{f}_2(0)}{2} - \frac{\dot{f}_2(0)}{2}. \quad (125)$$

(121) follows by taking the derivative of the expression in (119) with respect to θ . (122) is obtained by simplifying (121). We apply L'Hopital's rule on (122) to get (123) and further simplify it in (124). Finally, we obtain (125), which we used to derive (50) by inserting (112) and (116) into (125).

H. Proof of Theorem 10:

When the arrival rate is fixed, the following equality holds:

$$r_{\text{avg}}^*(\text{SNR}, \theta) = C_E(\text{SNR}, \theta). \quad (126)$$

Therefore, in formulas (54), (56), we can use $\dot{C}_E(0)$ and $\ddot{C}_E(0)$ instead of $\dot{r}_{\text{avg}}^*(0)/m$ and $\ddot{r}_{\text{avg}}^*(0)/m$ respectively, where we have defined $C_E(\text{SNR}, \theta) = C_E(\text{SNR}, \theta)/m$ as the normalized effective capacity. Minimum energy per bit and wideband slope becomes

$$\frac{E_b}{N_{0 \min}} = \frac{1}{\dot{C}_E(0)}, \quad (127)$$

and

$$S_0 = -\frac{2(\dot{C}_E(0))^2}{\ddot{C}_E(0)} \log_e 2. \quad (128)$$

Thus, we only need to obtain the first and second derivatives of $C_E(\text{SNR}, \theta)$ with respect to SNR at SNR = 0 to determine the minimum energy per bit and wideband slope. We first express the effective capacity given in (13) as

$$C_E(\text{SNR}) = -\frac{1}{\theta m} \log_e g(\text{SNR}) \quad (129)$$

where we have defined

$$g(\text{SNR}) = \mathbb{E} \left\{ e^{-\frac{\theta}{\log_e 2} \sum_{i=1}^m \log_e(1 + \text{SNR} z_i)} \right\}. \quad (130)$$

Now, the first and second derivatives of $C_E(\text{SNR})$ with respect to SNR are easily seen to be given by

$$\dot{C}_E(\text{SNR}) = -\frac{1}{\theta m} \frac{\dot{g}(\text{SNR})}{g(\text{SNR})}, \text{ and} \quad (131)$$

$$\ddot{C}_E(\text{SNR}) = -\frac{1}{\theta m} \frac{\ddot{g}(\text{SNR})g(\text{SNR}) - [\dot{g}(\text{SNR})]^2}{[g(\text{SNR})]^2}, \quad (132)$$

where \dot{g} and \ddot{g} denote the first and second derivatives of the function g with respect to SNR and can be expressed as

$$\dot{g}(\text{SNR}) = -\frac{\theta}{\log_e 2} \mathbb{E} \left\{ \sum_{i=1}^m \frac{z_i}{1 + \text{SNR} z_i} e^{-\frac{\theta}{m \log_e 2} \sum_{i=1}^m \log_e(1 + \text{SNR} z_i)} \right\} \quad (133)$$

and

$$\begin{aligned} \ddot{g}(\text{SNR}) &= \frac{\theta}{\log_e 2} \mathbb{E} \left\{ \left(\sum_{i=1}^m \frac{z_i^2}{(1 + \text{SNR} z_i)^2} + \frac{\theta}{\log_e 2} \sum_{i,j=1}^m \frac{z_i z_j}{(1 + \text{SNR} z_i)(1 + \text{SNR} z_j)} \right) \right. \\ &\quad \left. \times e^{-\frac{\theta}{\log_e 2} \sum_{i=1}^m \log_e(1 + \text{SNR} z_i)} \right\}. \end{aligned} \quad (134)$$

Then, at SNR = 0, we have

$$\dot{C}_E(0) = \frac{\sum_{i=1}^m \mathbb{E}\{z_i\}}{m \log_e 2} = \frac{\mathbb{E}\{z\}}{\log_e 2} \quad (135)$$

and

$$\begin{aligned} \ddot{C}_E(0) &= -\frac{\theta \sum_{i=1}^m \sum_{j=1}^m \text{cov}\{z_i, z_j\} + \log_e 2 \sum_{i=1}^m \mathbb{E}\{z_i^2\}}{m(\log_e 2)^2} \\ &= -\frac{\theta \sum_{i,j=1}^m \sum_{j=1}^m \text{cov}\{z_i, z_j\} + m \log_e 2 \mathbb{E}\{z^2\}}{m(\log_e 2)^2} \end{aligned} \quad (136)$$

where we have used the facts that $\sum_{i=1}^m \mathbb{E}\{z_i\} = m\mathbb{E}\{z\}$ and $\sum_{i=1}^m \mathbb{E}\{z_i^2\} = m\mathbb{E}\{z^2\}$ due to our assumption that the fading coefficients and therefore $\{z_i\}$'s are identically distributed.

Plugging the expressions in (135) and (136) into those in (127) and (128), we readily obtain the minimum energy per bit and wideband slope expressions in (57) and (58). ■

I. Proof of Theorem 11:

To show the result, we need to obtain the first and second derivatives of $r_{\text{avg}}^*(\text{SNR})$. We first express the maximum average arrival rate in (34) as

$$\begin{aligned} r_{\text{avg}}^*(\text{SNR}, \theta) &= \frac{P_{\text{ON}}}{\theta} \left[\log_e(1 - p_{11}g(\text{SNR})) - \log_e(g(\text{SNR})) \right. \\ &\quad \left. - \log_e((1 - p_{11} - p_{22})g(\text{SNR}) + p_{22}) \right] \end{aligned} \quad (137)$$

where we have used the definition that $e^{\theta C_E(\text{SNR}, \theta)} = \frac{1}{g(\text{SNR})}$ with $g(\text{SNR})$ defined in (130). Taking the first derivative with respect to SNR, we obtain

$$\begin{aligned} \dot{r}_{\text{avg}}^*(\text{SNR}, \theta) &= \frac{P_{\text{ON}}}{\theta} \left[\frac{-p_{11}\dot{g}(\text{SNR})}{1 - p_{11}g(\text{SNR})} - \frac{\dot{g}(\text{SNR})}{g(\text{SNR})} \right. \\ &\quad \left. - \frac{(1 - p_{11} - p_{22})\dot{g}(\text{SNR})}{(1 - p_{11} - p_{22})g(\text{SNR}) + p_{22}} \right]. \end{aligned} \quad (138)$$

Next, we let $\text{SNR} \rightarrow 0$. Since the arrival rate $\lambda \rightarrow 0$ when $\text{SNR} \rightarrow 0$, the equality in (138) becomes

$$\dot{r}_{\text{avg}}^*(0, \theta) = \frac{\dot{g}(0)}{\theta} P_{\text{ON}} \left[-\frac{p_{11}}{1 - p_{11}} - 1 - \frac{1 - p_{11} - p_{22}}{1 - p_{11}} \right] \quad (139)$$

$$= -\frac{\dot{g}(0)}{\theta} = \frac{1}{\log_e 2} \sum_{i=1}^m \mathbb{E}\{z_i\} = \frac{m\mathbb{E}\{z\}}{\log_e 2} \quad (140)$$

where $P_{\text{ON}} = \frac{1-p_{11}}{2-p_{11}-p_{22}}$. Plugging the result in (140) into (54), we immediately obtain (59).

In order to find the wideband slope, we first determine the second derivative of the maximum average arrival rate with respect to SNR and then evaluate it at SNR = 0 as follows:

$$\begin{aligned} \ddot{r}_{\text{avg}}^*(0, \theta) &= \frac{\ddot{g}(0)}{\theta} P_{\text{ON}} \left[-\frac{p_{11}}{1 - p_{11}} - 1 - \frac{(1 - p_{11} - p_{22})}{1 - p_{11}} \right] \\ &\quad + \frac{[\dot{g}(0)]^2}{\theta} P_{\text{ON}} \left[-\frac{p_{11}^2}{(1 - p_{11})^2} + 1 + \frac{(1 - p_{11} - p_{22})^2}{(1 - p_{11})^2} \right] \\ &= -\frac{\ddot{g}(0)}{\theta} + (1 - \eta) \frac{[\dot{g}(0)]^2}{\theta}. \end{aligned} \quad (141)$$

(141) follows from the fact that $g(0) = 1$, and η is defined in (37). Finally, inserting (140) and (141) into (56), the wideband slope expression in (60) is readily obtained. ■

J. Proof of Theorem 12:

We differentiate the maximum average arrival rate expression in (42) with respect to SNR and obtain

$$\begin{aligned} \dot{r}_{\text{avg}}^*(\text{SNR}, \theta) &= P_{\text{ON}} \left\{ \frac{2\theta \dot{C}_E(\text{SNR}) C_E(\text{SNR}) + (\alpha + \beta) \dot{C}_E(\text{SNR})}{\theta C_E(\text{SNR}) + \alpha} \right. \\ &\quad \left. - \frac{[\theta C_E^2(\text{SNR}) + (\alpha + \beta) C_E(\text{SNR})] \theta \dot{C}_E(\text{SNR})}{(\theta C_E(\text{SNR}) + \alpha)^2} \right\}. \end{aligned} \quad (142)$$

As $\text{SNR} \rightarrow 0$, we can easily derive

$$\dot{r}_{\text{avg}}^*(0, \theta) = P_{\text{ON}} \frac{\alpha + \beta}{\alpha} \dot{C}_E(0) = \dot{C}_E(0) = \frac{m\mathbb{E}\{z\}}{\log_e 2} \quad (143)$$

where we use the facts that $C_E(0) = 0$ and $P_{\text{ON}} = \frac{\alpha}{\alpha + \beta}$. Plugging the result in (143) into (54), we immediately obtain (65).

In order to determine the wideband slope, we additionally take the second derivative of the maximum average arrival rate with respect to SNR and evaluate it at SNR = 0 as

$$\ddot{r}_{\text{avg}}^*(0, \theta) = \ddot{C}_E(0) - \frac{2\theta\beta}{\alpha(\alpha + \beta)} \left(\dot{C}_E(0) \right)^2. \quad (144)$$

Now, inserting the results in (143) and (144) into (56) and using the formulations in (135) and (136), we obtain (66). ■

K. Proof of Theorem 13:

The proof is rather straightforward after realizing that $r_{\text{avg}}^*(\text{SNR}, \theta)$ of the MMPP source given in (48) is equal to the maximum average arrival rate of the Markov fluid source in (42) scaled with $\frac{\theta}{e^\theta - 1}$. Therefore, making use of the results in

(143) and (144), we can immediately express the first and second derivatives of $r_{\text{avg}}^*(\text{SNR}, \theta)$ at $\text{SNR} = 0$ as

$$\dot{r}_{\text{avg}}^*(0, \theta) = \frac{\theta \dot{C}_E(0)}{(e^\theta - 1)} = \frac{\theta m \mathbb{E}\{z\}}{(e^\theta - 1) \log_e 2}, \quad (145)$$

$$\ddot{r}_{\text{avg}}^*(0, \theta) = \frac{\theta}{(e^\theta - 1)} \left[\ddot{C}_E(0) - \frac{2\theta\beta}{\alpha(\alpha + \beta)} \left(\dot{C}_E(0) \right)^2 \right]. \quad (146)$$

Then, the expressions in (70) and (71) are obtained by plugging (145) and (146) into (54) and (56). ■

REFERENCES

- [1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018. Available online at <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-white-paper/c11-520862.pdf>
- [2] S. Tanwir and H. Perros, "A survey of VBR video traffic models," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 4, pp. 1778-1802, 2013.
- [3] D. Feng, C. Jiang, G. Lim, L. J. Cimini Jr., G. Feng, and G. Y. Li, "A survey of energy-efficient wireless communications," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, pp. 167-178, 2013.
- [4] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys & Tutorials*, vol. 13, no. 4, pp. 524-540, 2011.
- [5] R. L. Cruz, "A calculus for network delay, part I: Network elements in isolation," *IEEE Trans. Inf. Theory*, vol. 37, pp. 114-131, Jan. 1991.
- [6] R. L. Cruz, "A calculus for network delay, part II: Network analysis," *IEEE Trans. Inf. Theory*, vol. 37 pp. 132-141, Jan. 1991.
- [7] R. L. Cruz, "Quality of service guarantees in virtual circuit switched networks," *IEEE J. Select. Areas Commun.*, vol. 13 pp. 1048-1056, Aug. 1995.
- [8] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Auto. Control*, vol. 39, no. 5, pp. 913-931, May 1994.
- [9] C.-S. Chang, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol.13, pp. 1091-1100, Aug. 1995.
- [10] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," *Proc. of IEEE INFOCOM*, pp. 1001-1009, 1995.
- [11] C.-S. Chang, *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.
- [12] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. Springer-Verlag, 2008.
- [13] F. Kelly, "Notes on effective bandwidths," *Stochastic Networks: Theory and Applications*, Royal Statistical Society Lecture Notes Series, 4., pp: 141-168, Oxford University Press, 1996.
- [14] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. on Networking*, vol. 1, no. 3, pp. 329-343, Jun. 1993.
- [15] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. on Networking*, vol. 1, no.4, pp. 424-428, Aug. 1993.
- [16] C. Courcoubetis and R. Weber, "Effective bandwidth for stationary sources," *Prob. Eng. Inf. Sci.*, 9, pp. 285-296, 1995.
- [17] A. L. Stolyar, "Large deviations of queues sharing a randomly time-varying server," *Queueing Systems*, vol. 59, pp. 1-35, 2008.
- [18] V.J. Venkataramanan and X. Lin, "On wireless scheduling algorithms for minimizing the queue overflow probability," *IEEE/ACM Trans. Networking*, vol. 18, no. 3, pp. 788-801, June 2010.
- [19] B. Sadiq and G. de Veciana, "Large deviations sum-queue optimality of a radial sum-rate monotone opportunistic scheduler," *IEEE Trans. Inform. Theory*, vol. 56, no. 7, pp. 3395-3412, July 2010.
- [20] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol.2, no. 4, pp. 630-643, July 2003.
- [21] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assurance," *IEEE Trans. Veh. Technol.*, vol.53, no.5, pp. 1547-1557, Sep. 2004.
- [22] J. Tang and X. Zhang, "Quality-of-Service Driven Power and Rate Adaptation over Wireless Links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058-3068, Aug. 2007.
- [23] L. Liu, P. Parag, J. Tang, W. Y. Chen and J. F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Trans. Inform. Theory*, vol. 53, no. 5, pp. 1767-1777, May 2007.
- [24] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 2318-2328, June 2008.
- [25] S. Shakkottai "Effective capacity and QoS for wireless scheduling," *IEEE Trans. Automatic Control*, vol.53, no. 3, pp. 749-761, April 2008.
- [26] L. Musavian and S. Aissa, "Effective capacity of delay-constrained cognitive radio in Nakagami fading channels," *IEEE Trans. Wireless Commun.* vol. 9, no. 3, pp. 1054-1062, March 2010.
- [27] S. Akin and M. C. Gursoy, "Effective capacity analysis of cognitive radio channels for quality of service provisioning," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3354-3364, Nov. 2010.
- [28] M. C. Gursoy, "MIMO wireless communications under statistical queueing constraints," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 5897-5917, Sept. 2011.
- [29] D. Qiao, M. C. Gursoy, and S. Velipasalar, "Transmission strategies in multiple access fading channels with statistical QoS constraints," *IEEE Trans. Inform. Theory*, vol. 58, no. 3, pp. 1578-1593, March 2012.
- [30] B. Soret, M. C. Aguayo-Torres, and J. T. Entrambasaguas, "Capacity with explicit delay guarantees for general sources over correlated Rayleigh channel," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 1901-1911, June 2010.
- [31] Q. Du and X. Zhang, "Statistical QoS provisioning for wireless unicast/multicast of multi-layer video streams," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 420-433, Apr. 2010.
- [32] M. C. Gursoy, D. Qiao, and S. Velipasalar, "Analysis of energy efficiency in fading channels under QoS constraints," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4252-4263, Aug. 2009.
- [33] A. Helmy, L. Musavian, and T. Le-Ngoc, "Energy-efficient power adaptation over a frequency-selective fading channel with delay and power constraints," *IEEE Trans. Wireless Commun.* vol. 12, no. 9, pp. 4529-4541, Sept. 2013.
- [34] G. Ru, H. Li, L. Liu, Z. Hu, and Y. Gan, "Energy efficiency of hybrid cellular with heterogeneous QoS provisions," *IEEE Commun. Letters*, vol. 18, no. 6, pp. 1003-1006, June 2014.
- [35] L. Liu, Y. Yang, J. F. Chamberland, and J. Zhang, "Energy-efficient power allocation for delay-sensitive multimedia traffic over wireless systems," *IEEE Trans. Vehicular Technol.*, vol. 63, no. 5, pp. 2038-2047, June 2014.
- [36] M. Ozmen, M. C. Gursoy, "Impact of channel and source variations on the energy efficiency under QoS constraints," *IEEE ISIT*, pp. 806-810, 2012.
- [37] A. Lozano, A. M. Tulino, and S. Verdú, "High-SNR power offset in multiantenna communication," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4134-4151, Dec. 2005.
- [38] J.-W. So, "Performance analysis of VoIP services in the IEEE 802.16e OFDMA system with inband signaling," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1876-1886, May 2008.
- [39] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," *Proc. of the IEEE INFOCOM*, 1995.
- [40] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [41] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inform. Theory*, vol. 48, no. 6 pp. 1319-1343. Jun. 2002.